Design of a Recommender System for Participatory Media Built on a Tetherless Communication Infrastructure

by

Aaditeshwar Seth

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2008

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

We address the challenge of providing *low-cost*, *universal* access of *useful* information to people in different parts of the globe. We achieve this by following two strategies. First, we focus on the delivery of information through computerized devices and prototype new methods for making that delivery possible in a secure, low-cost, and universal manner. Second, we focus on the use of participatory media, such as blogs, in the context of news related content, and develop methods to recommend useful information that will be of interest to users. To achieve the first goal, we have designed a low-cost wireless system for Internet access in rural areas, and a smartphone-based system for the opportunistic use of WiFi connectivity to reduce the cost of data transfer on multi-NIC mobile devices. Included is a methodology for secure communication using identity based cryptography. For the second goal of identifying useful information, we make use of sociological theories regarding social networks in mass-media to develop a model of how participatory media can offer users effective news-related information. We then use this model to design a recommender system for participatory media content that pushes useful information to people in a personalized fashion. Our algorithms provide an order of magnitude better performance in terms of recommendation accuracy than other state-of-the-art recommender systems.

Our work provides some fundamental insights into the design of low-cost communication systems and the provision of useful messages to users in participatory media through a multi-disciplinary approach. The result is a framework that efficiently and effectively delivers information to people in remote corners of the world.

Acknowledgements

This thesis is dedicated to my parents. To my mother and hero, for giving me the strength of will to pursue what I believed. To my grandmother and teacher, for giving me the valuable lesson that the only way to understand mathematics is to picture what the numbers and formulae want to say. And to my grandfather and idol, for giving me strength of character to lead a righteous life. Without them and all the sacrifices they made, I wouldn't be the same person.

The thesis wouldn't have been possible without my supervisors, Keshav and Robin. Keshav has been more to me than just an adviser in technical matters. A friend, a mentor, and a guide for life, he gave me the flexibility to pursue whatever I felt was appropriate, provided me with continuous guidance even beyond his areas of interest to help me work efficiently and remain focused, and gave me the ability to form a vision to make the world a better place. He is truly a great person, both professionally and in character. Robin stepped in at just the right time to work with me, and without her continuous encouragement, inspiring dedication, and organized approach to research, I couldn't have completed this thesis. She is the most sincere teacher and guide I have come across, and I consider myself extremely lucky to have gotten her attention. Together, Keshav and Robin provided me with the perfect combination of advisers that I could have ever hoped to have as a graduate student.

But it's been a long road. All the work was doable only with feedback and collaboration from my co-authors and lab-mates, including Jie, Nabeel, Earl, Hossein, Rowena, Sumair, Usman, Matei, Shimin, Sonia, David, Omer, Darcy, and Majid. They provided an electric environment to think and debate matters such as what I have never experienced elsewhere. Nabeel and I have also closely shared much of our experience as PhD students – the contradictions, the disappointments at innumerable paper rejections, the sporadic joyous moments at papers acceptances – probably he is the only one who understands what made this road really long, and gave me company all along.

My committee members have been an indispensable source of support to me. The warmth and suggestions I got from Peter, Abby, and Urs, encouraged me to persevere and continuously strive to improve my work. And Alex and Paul, who have been in my committee since the beginning, rather than having dismissed my goals as foolhardy, helped de-construct and put them back together into more reasonable objectives.

I am also very thankful to Trevor, Ronaldo, and others from CSCF, and to Gail, Margaret, and Jessica from administration, all of whom played a very important role by taking prompt action for my requests for resources and repairs, without which I wouldn't have been able to work efficiently.

Over the last almost-five enriching years at Waterloo, so many other people and things have been a part of this experience as well, that it is hard to choose and name only a few. My friends and the time we spent together will forever remain precious to me. The climbing sessions with James followed by long introspections about life, the crazy movies that Nabeel made me watch, the stimulating discussions with Neeraj and appetizing meals by Jyotsna, and the fun moments with Shaweta punctuated with childish skirmishes were all essential to my work. My roommates were the people closest to me throughout, being there for me when I needed them the most. Srinath has been a part of so many activities with me, right from debating the nature of 42 on Saturday mornings in Tim Horton's to long drives all across "Amerika", that it is hard to believe how two people can share so many common interests. And Naveen is the sweetest person I know, selfless and a great friend and a wonderful cook, who has fed me so many wonderful dishes that I've almost forgotten my grandmother's cooking. Christine, Raiju, Bala, Abir, and others, all played a part in making this work possible. And during some of my lowest times, my buddies from Udai came forward and reinforced my belief that all was not wrong with the world – Nitin, Akhilesh, Aniket, Sarvagya, and so many more, I cannot thank you enough. I cannot also not mention my friends from the Outer's Club at Waterloo and some of my closest friends in India, together with whom I explored the world and its wonders that re-energized me to get back to work. The beautiful bike rides with Dalia, the literally meditative climbs with Ritwik, and the Himalayan adventures with Aditya, Mitesh, Harsh, and Vikrant, were all a crucial part of my PhD experience and life long friendships.

I cannot end without thanking Gary and Jose from the Knight Foundation, for giving me an opportunity to translate this thesis into reality. Their stories have only inspired and encouraged me even more to move ahead, and to rest assured that I will always find the right people to join me on this odyssey.

Contents

List of Tables

List of Figures

Preface

Writing a book is a horrible, exhausting struggle, like a long bout of some painful illness. One would never undertake such a thing if one were not driven on by some demon whom one can neither resist nor understand –
George Orwell, Why I Write, 1946

This thesis is quite unconventional. PhD theses are normally very focused on some specific problem, while here we have tried to address many different and diverse problems. This does not represent a lack of focus on our part, but rather a continuous questioning exercise about the eventual goal and usefulness of our research efforts. Building new communication technologies raises many questions. What kind of information do people actually need? What kind of tools are appropriate for them to access and share this information? What benefit do various interaction tools and information ultimately bring for the people? What problems can technology solve, and which ones can it not solve? All these questions required a focus beyond some specific technology or algorithm, and drove us across disciplinary boundaries – forming models in our heads, breaking them down, building tools, changing them, putting them together into a coherent vision... But while we had some important success in answering these questions, and we introduced many technological innovations in the process, the questions have introduced even more questions, and trying to answer them is what we see as our goal over the coming years, easily beyond this thesis. This document has however been written from a technical standpoint, and the reader will find it easier to interpret it accordingly.

Given the broad range of topics covered in this thesis, we have tried to lead the reader gradually into the technical details, while still preserving the readability and flow in the writing. Many technical terms that may be unfamiliar to readers from different disciplines have been underlined in the text and explained in a glossary in Appendix A. We further suggest that readers should read Chapters 1 and 5 before reading the rest of the chapters. Chapter 1 outlines the overall objectives, while Chapter 5 gives an overview of concepts relevant to the recommendation algorithms. Together, the two chapters give a broad perspective on multiple facets of the thesis. The rest of the chapters can be read independently though we suggest reading Chapters 2, 3, and 4 in that order, and 6, 7, and 8 in that order. Chapter 9 finally summarizes the main findings of the thesis, and presents some discussions and avenues for future work.

Chapter 1

Introduction

The medium is the message – Marshall McLuhan, Understanding Media, 1964

As computers become more prevalent in everyday life, computer-mediated communication is fast becoming the avenue of choice for acquiring information. This can bring many benefits to human society for the following reasons:

- *Information growth*: Growth of literature follows growth in literacy as human knowledge expands and people from diverse backgrounds desire to gain access to this knowledge. Estimates suggest that there were only 150,000 literate people in the known world in 4000 BC [1]; the number now stands at 5.5 billion just 6000 years later. This growth in literacy has been followed by large volumes of information growth, as is clearly evident from over 950,000 books published each year currently, and 7 million Internet web pages added daily.

- *Computerized tools*: With such information growth rates, computerized tools become a critical way to index the information and to make it searchable to find useful information. Google, for example, claims to scan 3000 books per day and has indexed over 25 billion web pages so far [2].

- *Telecommunication*: Computerized tools have also made it easier and cheaper to gain access to information, making information available universally across geographical and political boundaries, on-demand, and across a wide variety of communication devices. Over 1.3 billion people access information on the Internet from their desktops and laptops. In addition, 3.3 billion cellphones are already operational in the world, and are projected to become Internet endpoints in the next few years [3].

- *Human development*: Human development follows naturally from information access. For instance, newspapers have helped create informed societies

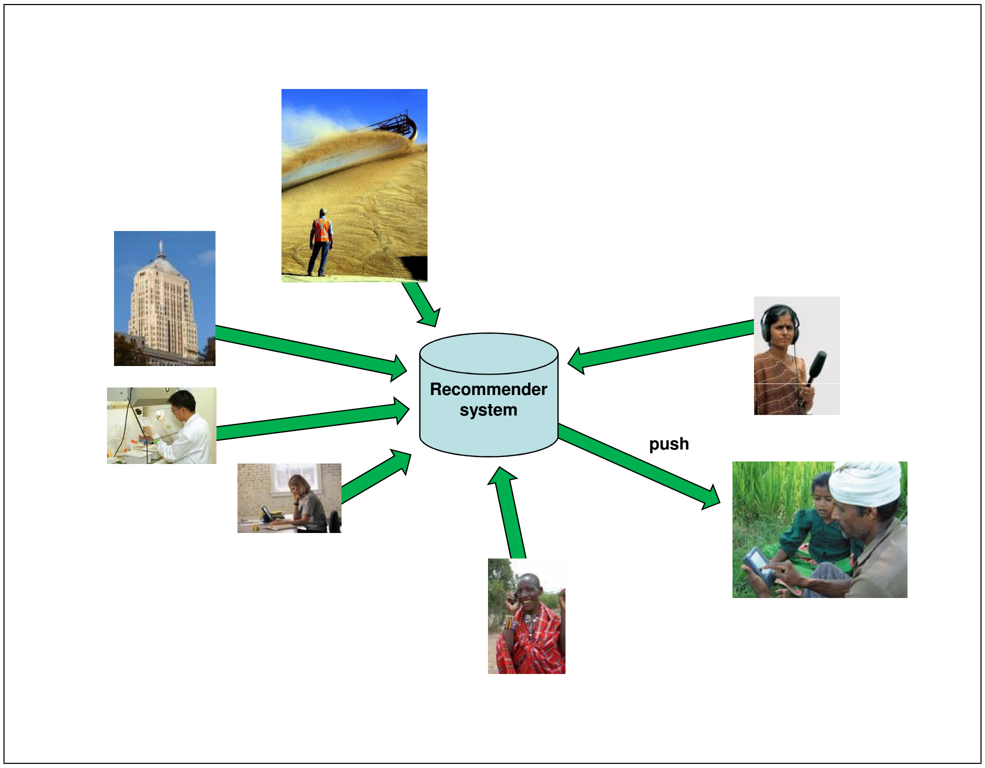


Figure 1.1: Use case

leading to more democratic participation of people in political processes [4]. Publication of technological developments related to health, engineering, and natural sciences in one part of the world have helped accelerate development in other parts [5]. Digitization of information and electronic communication only makes this easier, and evidence indeed shows a high correlation between increase in teledensity and the per capita GDP growth [6].

This leads us to conclude that in today's scenario computer mediated communication is the most important solution to help people find useful information and access it in a low-cost and universal manner, with the optimism that access to this information will improve human development. Our aim in this thesis is therefore to address the challenges that arise in providing low-cost, universal access of useful information to people in different parts of the globe. We next describe a use-case to illustrate the challenges that would need to be tackled to realize this vision.

Use case: The farmer as an information consumer

Consider the scenario shown in Fig.1.1 where a farmer living in a rural area in (for example) India is trying to decide his crop rotation pattern for the coming year. This can be a very complex matter. The farmer should ideally take into account (i) global commodity price projections which would affect the market price for his produce, (ii) expansion strategies of agro-technology corporations which would affect the cost structure and productivity for the farmer, (iii) internal politics of major wheat exporting countries, and (iv) technical factors such as soil conditions, irrigation requirements, and fertilizer usage which affects the sustainability of the rotation pattern. We assume for the purposes of this discussion that the farmer is literate and has access to a mobile device that can connect to the Internet at a low cost. We also assume that the device can access a **recommender system** on the Internet that automatically pushes to the farmer information about the issues listed above. Our vision is to build such a system and the supporting communication infrastructure. However, many challenges need to be overcome [1].

- Access to *low-cost communication* in rural areas is a problem. Some places in Kenya and Russia even today incur costs of up to $200 per month for dial-up Internet access [8]. Broadband penetration remains low; India has a mere 3.8 million broadband connections [9]. And the cost of data communication on cellular networks is also quite high, even in places with high cellphone penetration [10].

- Determining what information would be useful to the farmer is a challenge too. As we pointed out earlier, a wide variety of issues may need to be considered by the farmer, and the system should be capable of inferring what these issues are so that it can make recommendations to convey a *complete* picture to the farmer. Furthermore, the relative importance of various issues will depend upon the *context*: is the farmer in India or in Kenya? does the farmer directly sell his produce in a wholesale market or work on a contractual basis for some retail company? what is the risk absorbing capability of the farmer? etc. Therefore, the usefulness of information will depend upon both how complete it is and the context in which it is presented. In addition, since different farmers may have different contexts for information requirements, the recommender system should provide *personalization* on an individual basis.

  The inference of user context, the identification of relevant issues in order to give a complete picture, and personalized recommendation of information are all challenging problems to resolve the aim of designing a computerized system to determine appropriate information.

- The ease of publishing information on the Internet without any editorial checks by a formal agency makes it crucial to ensure that the information

[1] Note that so far we have only considered the aspect of providing useful information to the farmer. It may of course be argued that information is not the only problem, or even the most important problem, that farmers face today. Global food price inflation caused by improper governmental policies may be a bigger problem [7]. In this thesis, we however only concern ourselves with the role that information can play in solving these problems. Further, in this chapter, we will only discuss downstream information that should be pushed to the farmer. Later, in chapter 5, we will also discuss the upstream flow of information from the farmers to governmental agencies to enforce the formulation of appropriate policies, and show how our proposed system can facilitate this aspect as well.

recommended to the farmer is *credible*. However, credibility assessment is also a well known and hard problem.

- Fig.1.1 also shows community radio operators, local bloggers, and knowledgeable friends of the farmer who are likely to understand the context of the farmer. These entities help to filter out irrelevant and non-credible information for their audience, and present useful information in a *simplified* manner. The involvement of a diverse set of entities also helps ensure that *diverse* issues and opinions are presented to the farmer to convey a complete picture to him. The information produced by these entities is therefore of especially high usefulness for the farmer to help him cope with the complexity of the various issues that should be considered by him. We refer to this information as *participatory media content* since the people themselves participate in the information production process rather than institutions such as governments and newspaper agencies. However, with a greater number of entities participating in information production, the recommender system needs to *scale* to an even larger volume of information. For example, statistics suggest that over 1.6 million blog posts are written each day [11], but only a few would be useful to the farmer. Helping the farmer discover relevant blog posts from among this plethora of information makes the problem of discovering credible and useful information even harder.

- When comparing different crop rotation patterns, the farmer may make risk assessments based on the information recommended to him. We believe that the system should be able to *explain* to the farmer the extent to which this information addresses the important issues that need to be taken into account, that is, how complete is the information being recommended. Without doing so, the farmer may be at danger of not knowing that he does not know enough, which may incorrectly impact his risk assessments.

Each of the challenges described here are not trivially solvable. We list them again in the next section, and then give an overview of our solution approach.

Scope and goals

Given the importance of news media in society [4], we choose to focus on news information. Further, we place special emphasis on participatory media content related to news because participation by people in the information production process provides an avenue for information filtering and simplification to reduce the input for the recipient. We refer to participatory media content such as blog entries and discussion-forum postings as *messages*. Drawing insights from the use-case presented earlier, we aim to design a computerized recommender system with the following goals to push useful messages to users:

1. *Low-cost and universally accessible*: The system should be accessible universally and at a low-cost, to ensure democratic access and participation by the users.

2. *Personalized*: With the rapid rate of information production, personalized recommendation of useful messages should be provided. Personalization should take the *context* of the users into account, and the system should be capable of pushing messages to users in a *complete* manner.

3. *Support for credibility assessment*: Mechanisms should be designed to validate the credibility of participatory media content to ensure that credible messages are recommended to users.

4. *Scalable*: The system should be able to scale to millions of new messages published daily, and offer recommendations to hundreds of millions of users.

5. *Grounded in media theory*: Considerable research in the area of media and message effects has identified factors that should be considered to improve the effectiveness of news media [12, 13]. Similar to our definition, these factors define useful messages as messages that provide *simplification* and *diversity* of information, to help users gain clarity and unbiased viewpoints about various topics. As pointed out earlier, participatory media can help provide simplification and diversity because participants in the information production process understand the contexts of other people whom they know, and can hence simplify and provide a complete picture to them. However, current recommender systems do not seem to explicitly model the goal of recommending simplified and diverse information to users. For example, past studies on the search services of Google showed that its algorithms tend to bias results towards more popular websites and reduce diversity [14]; although this was challenged subsequently [15], it does indicate uncertainty in whether diversity is indeed assured in the results.

   Not only should diversity in the results be assured, *explanations* of the degree of diversity should also be provided to users so that they are aware of the degree of completeness in the information that is pushed to them. Explanations are also critical to system designers so that they know whether the information being recommended by their system is indeed useful and improves media effectiveness. Although we do not explicitly solve the problem of generating explanations as a part of this thesis, our intuition is that explanations will be easier to provide if the system design and recommendation algorithms are grounded in media theory from the very beginning. We therefore consider it worthwhile for insights from media research to be used to form the theoretical foundations of recommender systems for news media, so that their output can be *explained* in sociological terms.

We propose to meet these goals by taking a multi-disciplinary approach spanning the fields of computer networks, information retrieval, artificial intelligence,

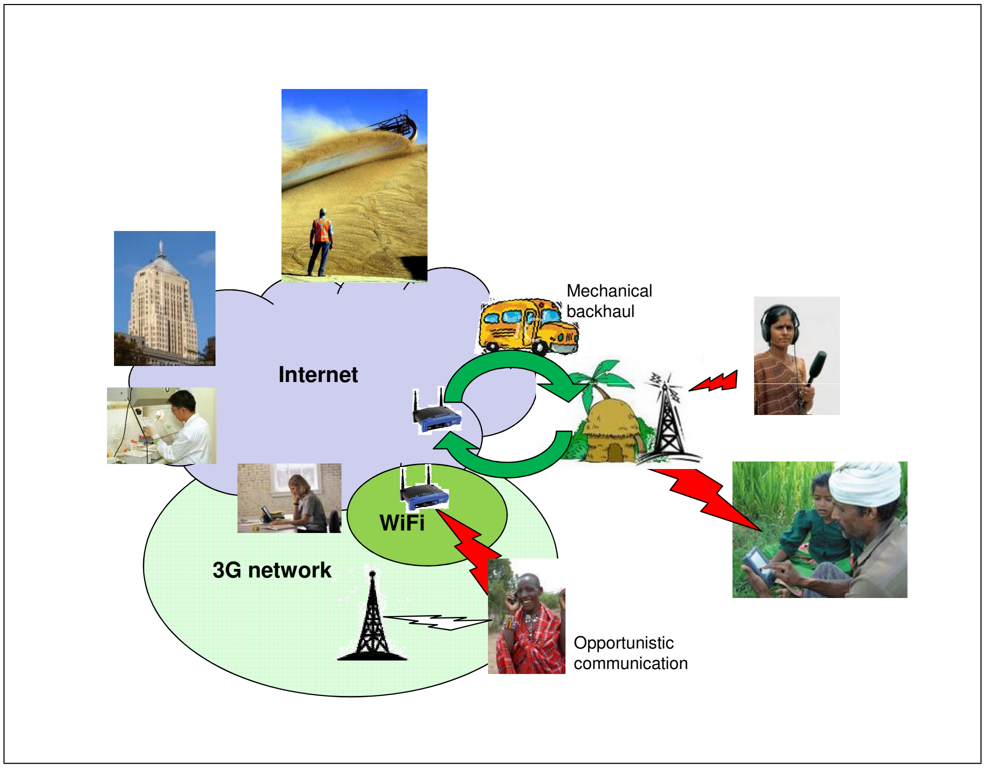


Figure 1.2: Delay tolerance reduces cost

information science, media research, and social network theory.

We will add that taking human factors into account in the design of information systems not only helps to simplify and improve the design of the system, but also helps understand the behavior of the system. This is important because technological systems tend to become the "extensions of man" and shape society for the years to come [16]; hence their sociological effects should be as well understood as possible [17]. An understanding of these effects should be considered a responsibility by system designers; this further justifies our use of a multi-disciplinary approach.

Solution approach

Our solution approach is based on two fundamental insights:

- *Delay tolerance reduces cost*: Many applications do not impose requirements on real time data delivery. This is true to a large extent for news related information – depending upon the criticality of a topic, users may be insensitive to whether they receive the news within an hour or a day or a week. Information such as opinions or reflections on past events tend to be even less sensitive to delay. Our insight is that delay tolerance can reduce communication costs. We explain this through the following scenarios:

- *Opportunistic communication on cellphones*: The cost of download and upload of data on cellular networks remains prohibitively high. WiFi presents a cheaper and often free alternative because of its use of the unlicensed spectrum. However, the spotty coverage provided by short-distance WiFi networks implies that connectivity will not be available all the time. We propose that WiFi networks should be used opportunistically when a user moves into and out of coverage as she walks or drives past WiFi hotspots. Thus, data may not be transferred instantaneously but will have to be buffered until the next connection opportunity. Delay tolerance implies that such opportunistic data transfer will not impact the usability of the system, yet will offer the benefits of lower communication costs.

- *Mechanical backhaul in rural areas*: The cost of providing communication in rural areas is mainly challenged by the last mile access. In the context of rural India, fiber PoPs are available in most small towns, but connectivity from these PoPs to remote villages is scarce [18]. This is because copper lines for DSL are not available in many areas, and wireless solutions such as long distance WiFi or WiMax are both expensive and complicated to set up [19]. Building on the pioneering work by Daknet [66], we propose that vehicles which regularly travel between villages and towns be mounted with a single-board-computer having WiFi and a large storage, and powered from the battery of the vehicle. The computer will wirelessly pick up data from mobile devices and kiosks in villages when the vehicle drives past them, and drop off the data when the vehicle enters a town and gains connectivity to the Internet through a WiFi hotspot. Deployment costs are drastically reduced because this "mechanical backhaul" system does not require any trenches to be dug or towers to be erected; even the cost of computers mounted on vehicles is amortized across all villages that a single vehicle visits. Delay tolerance again implies that the usability of the system will not be severely impacted, and yet the communication costs will be reduced.

Fig.1.2 illustrates the insight, based on which we will show in this thesis how the communication systems we propose can help to provide low-cost and universal access to information.

- *Social networks reveal context*: People are embedded in real-world social networks of ties between friends, acquaintances, family members, etc. This is shown in Fig.1.3. It has been noticed that social networks tend to be organized in clusters of people with a high density of ties between people within a cluster, but a sparse density across clusters [94]. This happens because people in the same cluster tend to know each other well and share a similar context in terms of income status, age, gender, geographical location, ethnicity, etc [95].

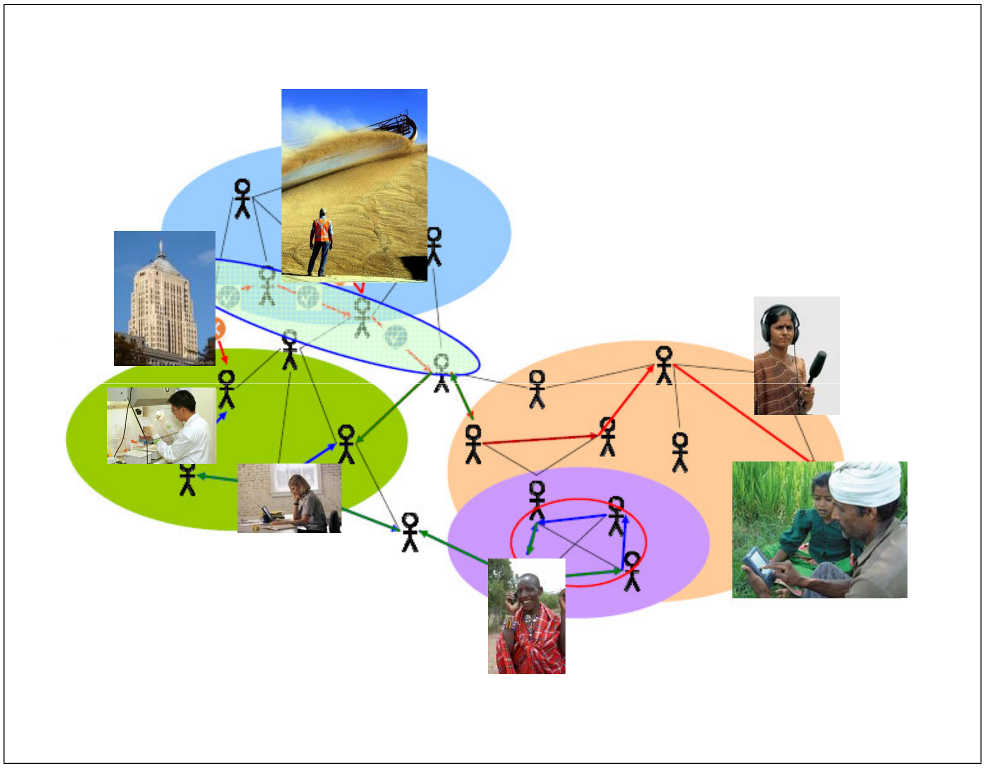


Figure 1.3: Social networks reveal context

Our insight is that information about the social networks of people reveals the contexts in which they are embedded. Interestingly, social network information is now becoming available through APIs such as those from Facebook (*www.facebook.com*) and the OpenSocial consortium (*code.google.com/apis/opensocial/*), and from email networks inferred by aggregating the address-books of users [22]. We will show in this thesis that we can in fact use this social network information to mathematically infer the contexts of people, and can use the models to improve the quality of recommendations in a personalized and context-sensitive manner. This becomes possible because we do not consider messages in alienation from their authors and recipients; rather, knowledge of the underlying social network relationships between an author and recipient reveals how well the author understands the context of the recipient [97]. This approach is used to determine whether a message will be simple enough for the recipient to understand, and whether the message will provide more complete information. The same approach in fact also helps to build models to assess the credibility of information, to improve the scalability of the system, and to explain the behavior of the system in sociological terms.

Both these insights together help us meet the goal of providing low-cost universal access to useful information for people in different parts of the globe. We are able to show in this thesis that our multi-disciplinary approach enables important distinctions from solutions offered by other researchers, and produces results for users that are in better agreement with their interests. The following contributions are made in this thesis:

- Design, prototype implementation, and evaluation of the Opportunistic Communication Management Protocol (OCMP) for opportunistic communication on mobile devices [184, 185].
- Design and deployment of the KioskNet project using mechanical backhaul based communication for Internet connectivity in rural areas [187].
- Design of a security protocol for KioskNet [188, 189].
- Definition of the theoretical constructs of context and completeness of messages, and validation of their relationship to the social network of message authors and recipients [191, 196].
- The use of social networks, and the constructs of context and completeness, to design and explore personalized Bayesian models for accurate recommendation and credibility-assessment of messages for users [192–194].
- Design of a distributed system to implement the proposed algorithms for personalized recommendation of participatory media content in a scalable manner [190, 195].

We are hopeful that our proposal for computer mediated communication will enable people to create and share information more effectively, and will lead to improvements in human development. The medium may also well indeed turn out to be the message, by enabling people to gain more awareness about each other, leading to a more intelligent society in which humans are mutually respectful of each other.

Outline

This thesis is structured as follows. Chapters 2 and 3 describe our work to provide low-cost Internet access on cellphones and in rural areas respectively. Chapter 4 introduces the critical factor of information security and presents our solution for ensuring security within the proposed framework for delivering low-cost access. Chapter 5 introduces participatory media in more detail, and outlines our approach of using social network information to design recommendation algorithms. Chapter 6 uses social network theory to propose and validate a communication model of user behavior for the production and consumption of participatory media content. This model is used in Chapter 7 to develop machine learning techniques that can classify, on a personalized basis, the credibility and usefulness of messages for users, and hence determine which messages to recommend to them. Chapter 8 describes the overall system architecture constituted of the various communication and recommendation components described in previous chapters, and presents additional studies to demonstrate the scalability of the proposed system. Finally, chapter 9 presents a discussion of the main contributions of our research and proposals for future work.

Chapter 2

Opportunistic Communication Management Protocol

When minds interact, new ideas emerge –
J.C.R. Licklider and Robert W. Taylor, The Computer as a Communication Device, 1968

In this chapter, we describe our solution to reduce the cost of data transfer to and from mobile devices. We do this by making opportunistic use of **WiFi hotspots** to connect to the Internet because WiFi (802.11) provides cheaper and faster connectivity than the cellular interface. We describe the rationale behind our approach in the next section, design goals in Section 2.2, design overview in Section 2.3, and a detailed description of the system architecture in Section 2.4. This is followed by a few implementation details in Section 2.5, an evaluation of the system in Section 2.6, and a discussion of related work in Section 2.7.

2.1 Multi-NIC devices and delay tolerance

The past few years have seen an explosive growth in the number of mobile devices such as cellphones, PDAs, and laptop computers. These devices have multiple NICs (network interface cards) and can use a variety of wireless access technologies ranging from wide-area technologies such as **GPRS, EDGE, CDMA 1xRTT, and EV-DO**, to local-area technologies such as 802.11 and short-range technologies such as **Bluetooth**. These wireless technologies differ from each other on a number of parameters, summarized in Table 1 [10, 24, 25].

In general, short-distance wireless technologies are better than long-distance technologies on parameters of data rate, power consumption per bit, and monetary cost:

- *Data rate*: Any wireless access technology must make a difficult tradeoff between the coverage area of an access point and the capacity available to a user in that access point's coverage area. Thus, long distance technologies such as EDGE have a range of a few kilometers, but support a lower data rate than short distance technologies such as WiFi, that have coverage areas of only a couple of hundred meters.

- *Power consumption*: The average power consumption of a cellular radio such as EvDO is of the same order as that of a WiFi radio, despite the larger communication distances, because cellular technologies use protocols for fine-grained open- and closed-loop power control that enable efficient channel management. However, due to the lower data rate of EvDO, the power consumption per bit of WiFi is much lower than EvDO.

- *Monetary cost*: The monetary cost of using an access network depends on the pricing policy of the network provider. It is practically zero in the case of private Bluetooth networks or free public access WiFi hot-spots operating on unlicensed spectrum, but can be very expensive (up to \$25/MB) for cellular data access plans.

However, along with these benefits, a different set of problems arise with the use of short distance technologies. For example, only a few CDMA base stations are sufficient to provide coverage to a large geographical area, but almost a hundred times the number of WiFi access points are required to cover the same area: this greatly increases the management overhead. Connection management for short distance technologies also becomes complicated because small coverage areas imply frequent disconnections and IP address changes during mobility.

To sum up, we believe that no single wireless access technology can be expected to provide *universal* high-bandwidth, power-efficient, and low-cost coverage. This points to the need for intelligent switching among multiple wireless interfaces, such

Table 2.1: Comparison of wireless technologies*

	Bluetooth	WiFi	GPRS	EvDO
Downlink data-rate (Mbps)	2.1	54	80Kbps	1.8
Coverage-radius (m)	∼ 10	∼ 100	∼ 1000	∼ 1000
Tx-power (mW)	130	1400	1250	3500
Rx-power (mW)	100	1150	600	1500
Energy(mJ)/Megabit**	47.6	21.3	7500	833
Cost/month	0	< \$20	\$30	\$80

* These are representative values only. Actual values can differ based on the modulation scheme used, distance from the base station, etc.

** Energy/Mbits in mW/Mbps = mJ/Megabit.

that users can opportunistically use one or more wireless networks to increase their overall communication capacity, power efficiency, and cost effectiveness [44–46].

Delay tolerant applications

In parallel, we observe that for many non-interactive applications, users can tolerate data delivery delays on the order of minutes, hours, or even days. Examples include non time-critical applications such as personal email, mobile blog uploads, download of media clips, etc. This flexibility in data transfer delays can be used to optimize communication overhead in novel ways. For example, data transfer can be intentionally delayed such that cellular networks are not utilized, and instead free WiFi or Bluetooth networks are opportunistically used whenever they become available. This can reduce the monetary cost per bit. It can also reduce the battery consumption on mobile devices because faster WiFi networks used for short bursts of time consume less power than slower cellular networks to transfer the same amount of data [47].

Data transfer may be delayed not just on mobile devices, but also within the infrastructure. This becomes useful in situations when the wireless link supports data rates much higher than the backhaul connection from a WiFi or WiMax access point. This is a typical scenario because most backhaul connections use Digital Subscriber Lines (DSL) that provide up to 5 Mbps, whereas 802.11g wireless links can support up to 54 Mbps. In such cases, WiFi access points can be enhanced with large local storage buffers. This will allow data transfer on the fast wireless link to proceed at the full data rate of the wireless technology: the data is buffered at access points and transmitted in batches over the slower backhaul link.

Summarizing this discussion, provision of delay tolerance and opportunistic connectivity on multi-NIC mobile devices can be used to reduce the cost of communication without significantly impacting the usability of non-interactive applications. Consider for example a user who has an email with photo attachments to send from her mobile device but prefers not to use the credit available on her expensive cellular plan. The user may, however, be willing to wait for a few hours (for example, 3 hours) in the hope of finding a free public hot-spot for sending this email. We would like to design a system that allows the user to express her connectivity preferences, but subsequently handles such a scenario transparently without requiring any user intervention in NIC selection or connection management. We next describe problems that need to be solved in order to build such a system.

2.2 Design goals

Fig.2.1 illustrates the above scenario for sending large photo attachments in a hypothetical WiFi and **3G** network layout. Here, a set of proxy (email cache) servers (marked P), located in different data centers in the Internet, are accessed by the

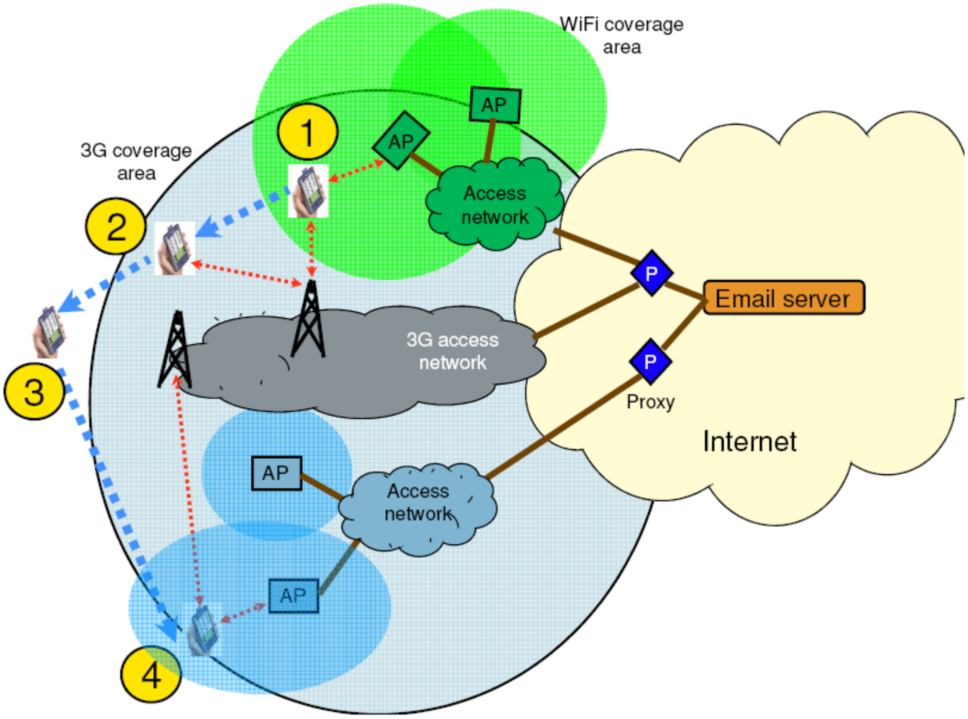


Figure 2.1: Opportunistic communication

mobile device. The device is initially at location (1) in overlapping 3G and WiFi wireless coverage areas. Because it can access the Internet using more than one wireless network, it needs to decide which network to use. The choice is complex, dictated by the dollar cost to use a wireless network, the power cost per bit, and the data rate, while simultaneously satisfying application requirements such as a 3 hour deadline for data delivery. How should these choices be resolved? Existing work on NIC selection only considers network properties such as cost and data rate to hierarchically rank networks with respect to each other [43, 44], and does not take delay based user preferences into account.

The device then moves to location (2), where it has only 3G coverage. How does it even know that it has left the WiFi coverage area? And how does it decide whether to use 3G, or to just wait for the next WiFi network? Should the user be involved in making this decision? This choice is equally complex, dictated by the user requirements, mobility schedule, and other network characteristics. Further, suppose the mobile decides to switch from WiFi to 3G because it predicts that it will not run into any WiFi network before the deadline. How can we hide this switch from the applications running on the mobile? Prior work to handle NIC switching or disconnections has only looked at small timescales at the network and transport layer [42, 45, 53]. Although systems such as [35, 41] do consider large timescale operations handled at the session layer, policy-based network selection using these systems has only considered the problem from a routing perspective, with a goal to minimize overall delays for data delivery [33]. To the best of our knowledge, the design of a general user-defined policy framework operating at the session layer has not been studied in the past.

Now suppose the device moves to location (3), where it has no coverage. At this location, the applications on the device have to deal with disconnection. How should this be hidden from applications, most of which are not disconnection-tolerant? Other systems handle this by constructing application specific plugins to hide disconnections from legacy applications [41], and we adopt a similar approach.

The device now moves to location (4), where it can access the Internet from the second WiFi access network. The device now has to decide whether this network is safe to use, and, if not, then use application-layer security mechanisms to protect privacy. Moreover, applications need to recover session state established in location (1) or (2) so that they do not start from scratch each time the mobile moves to a new location. Such scenarios have not been explored in existing work.

This example elicits the requirements that we enumerate below.

1. *Policy driven use of multiple networks*: Today's mobile devices can detect the availability of one or more access networks, but the interface selection is either completely manual, or based on a strict preferential order. This is clearly not suitable for opportunistic access. We would like a mobile device to have knowledge of factors such as the bandwidth and congestion state of the available wireless networks, energy-efficiency of the radios, and monetary costs of the networks. The device should then be able to use this knowledge to decide connectivity choices that satisfy the user requirements, and automatically connect to the appropriate networks. Thus, users should be able to specify policies such as data delivery deadlines, or bounds on monetary costs, or network preferences. Similarly, users should be able to specify application-independent preferences such as minimization of cost or power consumption. Whenever the mobile device is in the presence of multiple networks, it should be able to automatically select and connect to a suitable network in accordance with the user-defined policies.

2. *Support for legacy servers*: It is unlikely that a solution that requires changes at a server, for example the email server shown in Fig.2.1, will ever be deployed in practice. Consequently, we would like a solution that inter-operates with existing servers.

3. *Application session persistence across disconnections*: Consider that the mobile device has established a connection to a server using one of several network interfaces. Suppose the mobile decides to power itself off or switch to a different interface, and then reconnects to the same server with a different IP address. With existing systems, the server would be unable to recognize that the two connections correspond to a single ongoing data transfer session, and therefore would not be able to migrate persistent application state from the old connection to the new. For network access to be truly opportunistic

and seamless, an application should be able to exchange data with a server when changing network interfaces or even when faced with intermittent loss of connectivity. This requires the maintenance of persistent application state at both the server and the client so that data transfer can resume from the point where it stopped once connectivity is restored.

4. *Optimized network switching*: Consider that the user is walking or driving in a car, and the mobile device opportunistically connects to WiFi networks. Disconnections are likely to be unclean in such a scenario, and connections may terminate without an appropriate handshake that clears data in transit, or data sitting in local transport and link layer buffers. Although session layer persistence [26, 41, 42, 53] can ensure correctness that such data is not permanently lost, these protocols operate at timescales of the order of minutes (TCP timeouts); this can delay retransmission of data and result in large re-sequencing buffers at the receiver. It can also lead to the maintenance of redundant connection state at the proxy, because the proxy may wait for lengthy transport layer timeouts before tearing down the connections even though the mobile is already disconnected. Thus, we would like to have alternative mechanisms to better deal with network disconnections.

   Detecting and selecting a *good* network from among multiple networks also has room for optimization. Suppose the device is confronted with a choice of five to ten publicly accessible WiFi networks at the same time, which can be quite common in some urban settings. If the device connects to a few networks and probes the network quality to find a good network, valuable connection time can be wasted [48]. Again, optimization is required so that the mobile device can make quick decisions.

5. *Ease of application design and implementation*: Application designers who are familiar with the socket-bind-connect approach to writing distributed applications cannot deal well with systems where connections may fail arbitrarily, be resumed arbitrarily, and exhibit large variations in bandwidth depending on the currently available network. We would like to insulate application developers from these problems and provide them with a simple and intuitive communication interface.

6. *Support for buffered access points*: Access points enhanced with local persistent storage make better use of the faster wireless links because data is not bottle-necked at a slow backhaul link from the AP to an ISP. We would like to support such APs.

7. *Security*: The use of publicly accessible wireless networks clearly increases the security risks and requires safeguards from rogue access points, and eavesdroppers who can sniff wired or wireless communication. End-to-end security mechanisms such as SSL can be used in most cases, but support for delay tolerant infrastructure requires a rethinking of security solutions because the

end-to-end principles may not work efficiently when data is stored at intermediate locations. Furthermore, a chatty negotiation of public keys can be quite wasteful during an opportunistic connection. Therefore, we would like security solutions that can function during disconnection as well. We discuss our security solution later in chapter 4.

We address these goals by means of our system architecture and the Opportunistic Connection Management Protocol (OCMP). We present an overview of the system design in the next Section, deferring details of OCMP to Section 2.4.

Note that efficient opportunistic communication also requires solutions to issues such as 802.11 association delays, wireless losses due to interference and mobility, appropriate MAC rate adjustment, impact of lower layers on TCP and other transport layer implementations, etc. We defer an analysis of such factors to future work, and focus here only on the design of a policy-based architecture for opportunistic communication.

2.3 Design overview

Fig.2.1 also presents an overview of the system architecture. The main components are the content host, the proxies – which run OCMP servers, and the mobile host – which runs the OCMP client. We describe each component next.

2.3.1 Content host

At the right of Fig.2.1 is the content host, a server that either provides content such as video or stored voice to a mobile device, or receives uploads and content requests from the mobile. This represents popular web sites like youtube.com and flickr.com, or media servers that provide audio and video content. Content hosts reside in a data center at the core of the Internet. These servers are connected to a wired, high-capacity, and global IP core backbone. Existing content servers run legacy applications and do not support disconnection resilience or parallel transport connections over multiple networks for a single application session. We would like to provide a feasible path for supporting opportunistic communication *without* requiring modifications to legacy servers. We achieve this via the deployment of network-based proxies which are described next.

2.3.2 Proxy servers

Proxy servers (marked P) allow interworking between legacy servers and our protocols [27, 35, 41]. A proxy is located in the communication path between a mobile device and a content host. It serves as the termination point for the transport connections opened by the mobile host over multiple network interfaces. The proxy server hides multiple connections and disconnections from the content host. It can also provide fine-grained and application-specific connection management, as will be described in Section 2.4.

The proxy can either be provided by an Internet Service Provider, or by an enterprise on behalf of its employees. Proxies should be placed so that the round trip time from the proxy to the bulk of the mobile devices is as low as possible [184]. For example, cellular providers could keep the proxies adjacent to the **PDSNs** (Packet Data Serving Nodes) in CDMA or the **GGSNs** (Gateway GPRS Support Nodes) in GPRS networks, or on the backhaul point to the Internet core [185]. On the other hand, if a third party provides a proxy as a value-added service, it should place the proxy in a well-connected data center. We only require that the proxies have one or more globally-reachable public IP addresses, or dynamic DNS registrations.

The proxy acts as a store-and-forward agent for data downloads to a mobile device. A download starts with a mobile application initiating a data transfer request, e.g., an HTTP GET request. This request is intercepted by the OCMP client on the device and forwarded to the proxy, as in PCMP [41]. The OCMP server on the proxy supports an *application plug-in* that allows it to understand how to process application-specific data transfer requests. If the request is from an application supported by the proxy, the proxy processes the request and then uses legacy protocols to contact the content host on behalf of the mobile device. Thus the content host is completely shielded from all details of communication between the mobile and the proxy.

Once data is downloaded from the content host to the proxy, the proxy caches the data and looks for available connections to the mobile device. No such connection may be available at that time if the mobile device is temporarily disconnected from all access networks. If so, the proxy holds the downloaded data in persistent storage until the device reconnects. Alternatively, the proxy can use an out-of-band mechanism (e.g., an SMS message if the mobile device is a smart phone) to inform the mobile device about the availability of data. When the mobile device reconnects over one or more wireless networks, the proxy segments the application data into *bundles* (message units [35]) and routes the bundles over these connections. The policy-based algorithm for determining which bundles to transmit on what connections is negotiated in advance between the OCMP peers on the mobile client and the proxy.

When data is being uploaded from the mobile device to a content host (e.g., blog or picture uploads), the proxy receives bundles from a single application, potentially over multiple transport connections from the mobile, reassembles them into a single stream, opens a connection to the content host and forwards the data using application-specific protocols. Thus the OCMP client on the mobile host and the proxy implement analogous functions for segmentation and reassembly of application data as well as policy-based routing of application data segments.

We envision that the multi-connection state between a mobile and a proxy can be packaged and moved to a different proxy to allow a mobile to always use a *nearby* (**in terms of RTT**) proxy, greatly improving performance. Similarly, the state on a mobile device can be retained persistently across arbitrary periods of disconnection or power loss. The state can even be transferred to a different end-point like a home or office desktop, and unpackaged to recreate an operating state identical to the state on the mobile prior to disconnections. This can be used to provide semantics similar to that provided by Internet Suspend and Resume [38].

Note that we have not implemented a multiple-proxy system as yet. We do plan to implement inter-proxy state transfer in the future by arranging globally distributed OCMP proxies in an I3 overlay [34]. Mobile devices will be able to register themselves with the closest OCMP proxy (an I3 overlay node), and application state for the mobile device will be automatically moved between proxies whenever a new registration is made. Note that we expect OCMP proxies to have fairly large footprints [1], and therefore selection of the closest proxy need not be done on the order of every minute or even every hour.

2.3.3 OCMP

The proxy and a mobile may be connected by multiple heterogeneous wireless networks that differ in coverage, capacity, pricing, and availability. An OCMP server-side protocol running on the proxy coordinates access on these networks with an OCMP client-side protocol running on each mobile. OCMP defines a message format for encapsulating application data segments for session-level data reassembly. OCMP also supports control messages, i.e., messages that consist of only an OCMP header and an empty body. For example, control messages are exchanged between the OCMP client and the proxy to coordinate policies regarding selection of network connections as described in Section 2.4.8.

Unlike past work [42, 53], OCMP does not depend on TCP semantics of the underlying connections, as long as the transport layer provides end-to-end reliability and flow control. An underlying connection can be a standard or modified TCP/IP connection [31, 39] or can be a transport protocol optimized for wireless networks, such as erasure-coded UDP [29, 30]. OCMP can therefore exploit systems that compress and transcode data on wireless links to optimize bandwidth use [57].

Using OCMP, data can be striped on multiple connections in parallel, under fine-grained application control. Essentially, OCMP clients and servers choose the connection to be used for each application-level data unit based on pre-specified policies. Moreover, if a connection abruptly terminates, or even if *all* the connec-

[1] As of 2005, the Sprint cellular network was comprised of 12 distribution sites across USA. Each site catered to $\sim$ 100 base station controllers, and each such controller catered to $\sim$ 300 base station transmitters. Since we expect OCMP proxies to be located at the distribution sites, the footprint of a single proxy will be quite large.

tions terminate, the OCMP client and server gracefully recover from the failure, providing applications the illusion of seamless connectivity.

Besides working with the server-side, the OCMP client-side also has the additional responsibility of detecting network connections and disconnections. It uses application-specific policies to decide whether it should initiate a connection to the server side when a connection opportunity arises. It also has a notification mechanism to inform an application if there is any data that has arrived for it.

Connections between an OCMP client and OCMP server are always initiated by the client. A new transport layer connection is created each time the device connects on a new network and torn down when the device disconnects. Each network interface is associated with a single transport connection that is shared by all application data units assigned to that interface.

2.3.4 Mobile

The proxy identifies a mobile device (and all connections originating from it) by a globally unique identifier (which could be an **IMSI** or phone number). The GUID for a mobile device is common to all its transport connections, and serves several purposes. It decouples device addressing (a GUID) from routing (in terms of IP addresses), as in HIP [40]. This solves the problem of IP address changes due to mobility and/or disconnections. It also enables the proxy to maintain persistent data transfer state even when a mobile application uses parallel transport connections over different access networks.

The OCMP client on a mobile device provides two application interfaces. The first is meant for legacy applications that are designed in the socket-bind-connect paradigm. For such applications, application-specific data download requests are intercepted by the OCMP client and sent on a control connection to the proxy. The OCMP server layer in the proxy, acting on behalf of the client, initiates a connection to the server and downloads the data. It then transmits the data to the mobile device. The OCMP client layer on the mobile device reassembles the data before delivering it to the application. For data uploads, the proxy reassembles data received from the device and transmits it to the server.

We have also built a new application interface for disconnection- and delay-tolerant applications. It takes the form of a *communication directory*, which is a standard directory in the file system. An application writer drops a file into this directory, and is guaranteed that the file will appear at a corresponding directory on the destination at some point in the future. Policies can be associated with each directory, and can be defined through a configuration file for different classes of applications. This is described in more detail in Section 2.4.1. We have found that this API is both robust and easily understood by application developers.

We have developed a Java-based prototype for the system described above and evaluated its performance on laptops with diverse wireless access technologies such as 802.11b/g, CDMA 1xRTT, and GPRS EDGE. A detailed performance evaluation of our prototype is presented in Section 2.6.

2.3.5 Buffered Access Points

Whereas our solution will work with off-the-shelf WiFi or WiMax access points, we recommend enhancing these access points with store-and-forward infrastructure. This allows for better utilization of the wireless link capacities, which are typically much higher than backhaul bandwidths for public access networks.

The **DTNRG Delay Tolerant Networking** software [32] provides the required functionality. DTN is a style of networking that enables store-and-forward data transfer even if an end-to-end connection between two endpoints is not available. Such scenarios typically arise in challenged environments like inter-planetary communication, sensor networks, underwater networks, and connecting remote rural areas to the Internet. We have developed prototype access points running DTN using single-board computers from Soekris Corporation [56] fitted with 40 GB hard-disks. They interface with OCMP running on the mobile devices on the wireless network, and OCMP running on the proxies in the Internet on the wired network.

Such buffered access points are ideal for data uploads from mobile devices. Data can be opportunistically transferred to the intermediate access points at the highest data rates possible, and session level persistence ensures that the data is correctly reassembled and conveyed to the proxy.

Data downloads are more difficult because the proxy servers need to be aware of the next access-point that the mobile will connect to pre-cache downloaded data, and this depends on the mobility pattern of the users. We envision using network coding for redundancy and to compensate for incorrect mobility prediction by replicating data on multiple access points, such that if the mobile device is able to pick up a certain minimum amount of data, it can decode and reassemble the rest of the data. Although we have not addressed the problems of mobility prediction and optimal network coding in this chapter, our implementation is flexible to accommodate such features in the future.

2.3.6 Control channel

We observe that cellular networks are nearly ubiquitous, even though they may not support high data rates. We use this insight to assign a special role in OCMP to the cellular network, other than its regular use for data transfer. The cellular network in OCMP provides a *control plane*. For example, when mobile devices are confronted with a choice of multiple WiFi networks to choose from, they can use the control channel to query a centralized GIS (Geographic Information System) for the best-performing network at that time [54], or report back the performance status of different networks to keep the database updated in real time. Similarly, the control channel can be used for DHCP negotiations and pre-authentication to achieve small WiFi drive-thru association latencies. The control channel can also be used to send disconnection notifications for more efficient handling of in-transit or buffered data that is lost during unclean disconnections, and reduce the amount of redundant connection state being maintained.

2.3.7 Policy control

OCMP allows policy definitions at the application-specific and application- independent levels, and provides a framework in which algorithms can be implemented to demonstrate these policies. Any configuration parameters required by policies are assumed to reside in a policy module in OCMP, and can be queried to enforce policy at two key decision points.

1. *Which network to connect to*: When faced with a choice of multiple networks which may be of different types, decisions can be made based on the policy specifications.

2. *Which application data unit to schedule for data transmission*: When multiple applications compete for a common network resource, decisions have to be made based on the policy specifications.

OCMP is designed such that policy enforcement algorithms and policy configuration can be distributed among different entities. For example, policy configuration data may reside on the mobile device, but proxies may remotely query the policy module on the mobile device. If the enforcement algorithms are complex, then the algorithm evaluation can be performed in the infrastructure instead of using valuable battery and computation resources on the mobile device. The results of the algorithm can then be conveyed to the mobile device and enforced locally using a control channel.

In the current implementation, the policy enforcement algorithm to be used is specified as a startup parameter in OCMP. This algorithm has two methods: (a) *connect*, which is called each time a new network is detected, and returns *true* or $false$ to indicate whether to connect to this network or not, and (b) *getBundle*, which is called whenever the transport queue of an active network is vacant, and returns a *bundle* to be dispatched on this network, or returns *null*. We have implemented only a simple policy so far, described later in Section 2.4.8. However, any other kind of a policy enforcement algorithm can be used.

1. Policies can be modeled as utility functions, where the utility gained is greatest when application data is delivered immediately, and the utility decays with time, eventually going to zero after the deadline. An example utility function can be $U = a - bt$, where a and b are constants, and t is the time

difference between the time of data delivery and the time when the application started. Algorithms can now be designed to schedule applications such that the overall utility is maximized. This is the approach followed in [50], although the focus is on real-time applications where utilities can be based on round-trip latencies or loss rates or cost.

Application- independent utility functions can also be designed, so as to minimize power consumption and maximize the total application-specific utility simultaneously.

Note that the performance of such a scheduling algorithm may depend on whether or not it can successfully predict the expected mobility schedule of a user.

2. Policies can also be modeled as rules; for example, a policy may dictate to always use WiFi instead of EDGE, unless application deadlines are only an hour away. Such policy enforcement algorithms may require to implement a rules-engine, where rules can be specified in a policy definition language [49], along with suitable conflict resolution procedures.

Design of fast and optimal policy enforcement algorithms is an area of future work. This chapter describes the framework necessary to implement different policies.

2.4 System architecture

2.4.1 Client-side communication API

OCMP interacts with applications on the mobile client either by means of a *communication directory* or by intercepting socket calls made by legacy applications.

A communication directory is simply a directory in the file system that contains application data. To send data, an application creates a new file in the directory, or modifies an existing file. A *watcher* process periodically looks for modifications to the last modified time of the directory. If the modification time is more recent than the last time the directory was checked, the newly created or modified files are sent to the OCMP stack using the OCMP API.

The watcher also registers itself as a default plugin with OCMP, much like inetd. When called, it writes data to a file in the appropriate communication directory, and invokes an application-specific script to notify the application when its data has been received. The application can simply read this file to get its incoming data.

Each communication directory has two special files. The *config* file has application-specific configuration parameters. For example, for the blog-upload application, this is the username and password for the user. The config file also contains application-specific policies to control the interface(s) used for transferring data for that application. These policies are passed to OCMP by the watcher. The other special file is the *status* file. This file has one entry for each file in the communications directory and contains the status of that file. The status of a file can be, for example, *ready to send*, *partially sent*, or *sent*. An application that wants to know the status of a file's transfer can read the status file. This can be used, for example, to display progress in a user-friendly GUI.

The use of a communication directory simplifies application development. An application writer has to only create a send and receive directory and the associated config and status files. After that, all communication is achieved by writing files or reading files from the directory.

In addition to the communication directory, we support legacy Java applications by intercepting *socket* calls in the Java API. These calls are instead handled by OCMP, specifically by the application plugin associated with that application. OCMP infers the plugin associated with a socket call by looking at the destination port number as well as the first few bytes of the written data. We describe this in more detail in Section 2.4.4. After the interception, the remainder of the processing is identical as with the communication directory.

2.4.2 OCMP protocol stack

The OCMP client and server stacks that run on a mobile and on a proxy respectively are shown in Fig.2.2. On the client side, OCMP-aware applications interact with OCMP through a communication directory monitored by a *directory watcher*. Also, socket calls made by legacy applications are intercepted by OCMP, which redirects them to application-specific plugins.

We assume that applications or their associated plugins can categorize their communications into either a control or one or more data *streams*. For example, an email application may create a data stream for email bodies, and another stream for email attachments, where the delivery deadlines for attachments may be more relaxed than the deadlines for email bodies. The application control stream (shown by *App Ctrl* in Fig.2.2) provides an explicit control channel between the application plugin peers running on the mobile and proxy. For example, it is used to tell a receiver about the length of the bulk data sent on a data stream, or application parameters required by a peer plugin. It can also convey to the mobile the status of the data transfer between the plugin on the proxy and legacy servers.

Each application data stream is assigned to a *Storage manager* that segments/ reassembles the data into/from multiple bundles. It also stores the data in persistent local storage to deal with power loss on the device. Each data stream has an associated policy that is registered with the OCMP *policy module*. The policy module maintains a collection of preferences for data streams belonging to different

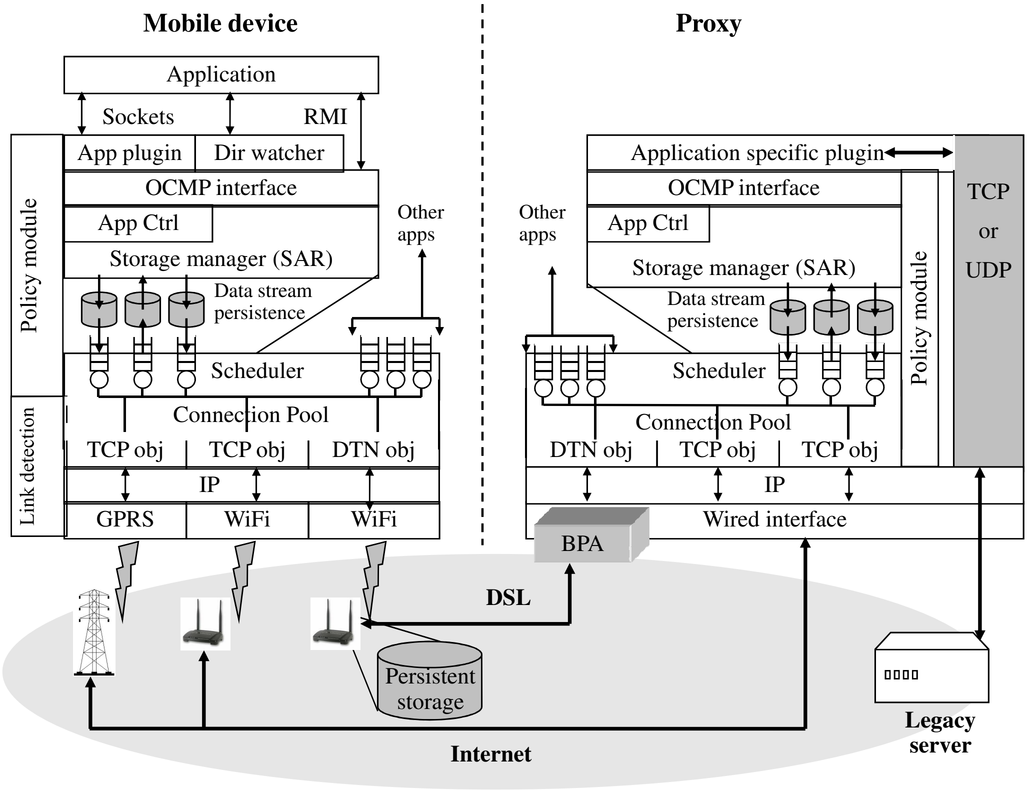


Figure 2.2: OCMP stack

applications. It also contains user-defined preferences. This set of user and application preferences are used by the OCMP *scheduler* to run policy enforcement algorithms to schedule application bundles on different interfaces.

The scheduler contains a *connection pool* object that maintains a list of active transport layer connections, one on each interface. Each connection is encapsulated in a *connection object*. The scheduler maintains a pre-fetch buffer for each data stream to minimize latencies in fetching data from the disk. It then selects bundles and sends them to one of the connection objects depending upon network availability and the user-specified policy. The scheduler can also decide what kind of a transport layer to use on each interface. The scheduler also may choose to associate on a network when a *link detection* module notifies it of new networks in range.

At a proxy, incoming bundles are processed by a symmetric stack and eventually handed to an application-specific plugin. These plugins can be loaded into OCMP dynamically on packet arrival. The plugin can then take application-specific actions to transfer the data to a legacy server. The plugin can also fetch data from a legacy server on behalf of an application and store it in data streams on the proxy. When a mobile opportunistically connects with the proxy, this data is sent to the mobile.

Note that the scheduler can be implemented either on the mobile or at the proxy. Computationally-efficient devices can run scheduling algorithms on the device itself and remotely set light-weight scheduling rules on the proxy. On the other hand, computationally-starved devices can shift schedule computation to the proxy, and rely on simple rules for bundle scheduling on the device. This flexibility is possible because the policy module that contains various user preferences can be queried remotely over the cellular control channel. Data on network status or mobility pattern can similarly be exchanged over the control channel to provide timely information to the scheduling algorithm. Note that in an area with no coverage, where the control channel is absent, no scheduling decisions are needed in the first place.

2.4.3 Session-level reliability

Due to the presence of a send buffer in the network stack, write calls that enqueue data into a non-empty send buffer return successfully, making an application think that the data was reliably delivered to the receiver, even though it might not be delivered at all if the connection closes prematurely! In this case, the data in the send buffer is actually lost after a connection termination is announced to the application. Similarly, on the receive side, bundles that have been acked by TCP, are not necessarily passed to the OCMP agent on an unclean disconnection. Hence, with any buffered protocol stack, there is always a possibility that data equal to the sum of the send and receive buffers is actually lost, even though the sending side believes the data to have been delivered successfully. For this reason, transport layer semantics are insufficient for reliable delivery, and a session level reliable data transfer protocol is needed to recover from lost data.

To avoid the overhead of a per-bundle ack or nak protocol, an OCMP sender keeps track of the order in which it transmitted data on each network interface, and retains, in persistent store, all data that might possibly get lost in transit. When a connection closure is detected, the receiver uses the control channel to inform the sender of the last sequence number bundle it successfully received on that connection. This allows the sender to infer the set of bundles that were not successfully received. The sender therefore queues them for transmission on a working interface, or marks them as undelivered for subsequent retransmission.

The ability for one end of a connection to inform the other of unclean connection termination on an alternate interface is a useful feature of OCMP. This is because we have found that in practice, one of the ends knows about a disconnection far sooner than the other. This technique allows both ends to reason correctly about the disconnection and to take corrective action. Typically, disconnections are due to wireless failures, which the mobile device finds out about much faster than the proxy. The mobile then sends a disconnection notification, along with the last sequence number it received on the WiFi interface, on the control channel. If the proxy was sending some data to the mobile on the WiFi interface, it can immediately retransmit the data sent on the failed interface after the last sequence number received by the mobile. The proxy responds to a disconnection message with a reply that carries the last sequence number it received on the failed interface. In case the mobile was uploading data, it can now retransmit everything it sent after the last sequence number that was received by the proxy. This allows us to quickly recover from a broken connection. We evaluate the performance of this technique in Section 2.6.

The discussion is illustrated in a sample scenario shown in Fig.2.3 for a mobile device that encounters intermittent WiFi connectivity and uses both WiFi and EDGE for data transfer. The protocol begins when the OCMP proxy notifies the mobile device that it has data waiting to be picked up by the device. We assume that these notifications can be sent through an out-of-band mechanism, such as SMS. When the mobile receives this notification, it asks the link detection module to raise an event whenever the device connects to a new network. Thus, when the mobile connects to a WiFi hotspot, the OCMP control layer decides to use TCP as a transport layer on WiFi to connect to the proxy. The connection is initiated through a control message, which first instantiates an OCMP *connection* entity for the mobile on the proxy if it did not exist already. The connection is then added into the connection pool. If the WiFi connection breaks uncleanly, the EDGE connection is used to send control messages to the proxy so that the proxy does not have to wait until a TCP timeout to detect the connection failure.

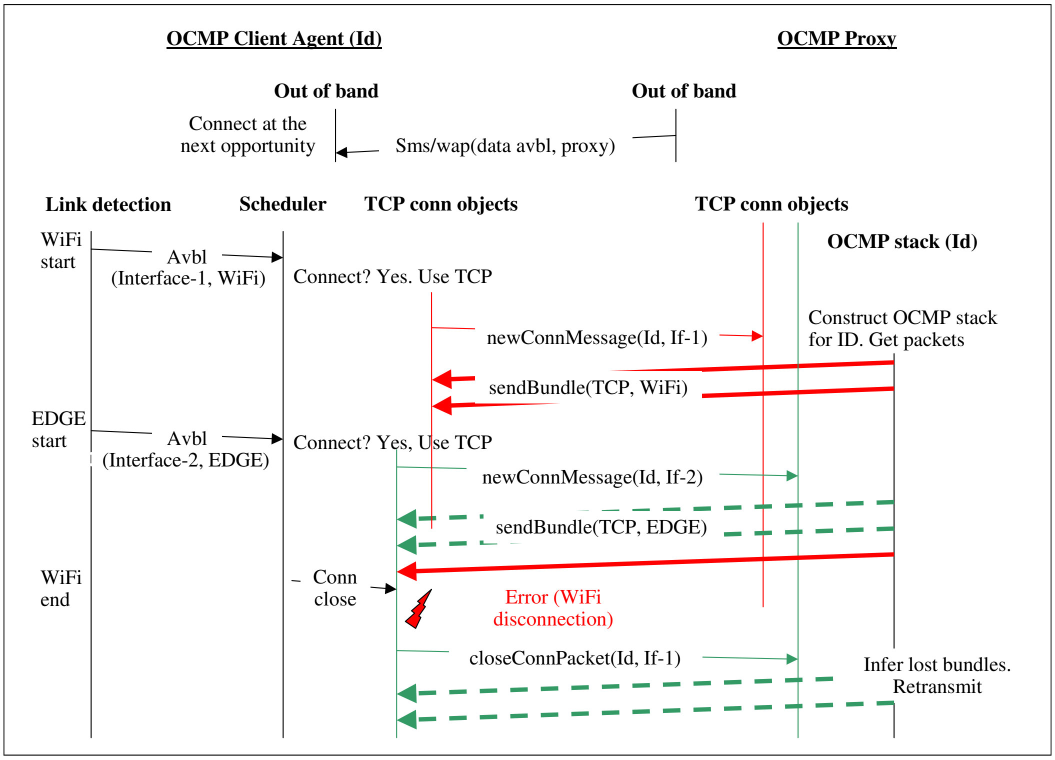


Figure 2.3: Control flow sequence diagram

2.4.4 Application-specific plugins

Both the OCMP client and the server support application-specific *plugins*. These short-lived code modules are invoked to carry out application-specific actions for each client-server interaction. All applications need a plugin at the proxy, and legacy applications need a plugin at the client end as well. For example, a legacy web browser request on a mobile is associated with an instance of a HTTP plugin both on the client on the proxy that initiates an HTTP GET on its behalf. The proxy-side plugin stores the results in persistent storage and communicates them to the client over opportunistic links. Other examples are a blog plugin to support upload from a mobile device to Blogger, and a Flickr plugin to upload a photograph to Flickr. Application plugins attempt to mask a mobile's disconnections from legacy applications either on the mobile or at the content host. Of course, long disconnections that last for hours or days cannot be masked, particularly from interactive applications. However, opportunistic communication is ideal for delay-tolerant applications, such as music downloads and email.

An instance of a plugin is created on the mobile if OCMP intercepts a socket call made by legacy applications. The destination port number or the first few bytes written into the socket are used to disambiguate applications, and a corresponding plugin object is created to handle the connections. Whenever a new plugin is created, or a new file is dropped into the *communication directory*, an application control message is also sent to the proxy to ask it to dynamically instantiate a peer plugin on the proxy. Application control messages have an application ID and application type field to uniquely identify the correct plugin and the type of the plugin. The plugin then collects *one-time* data from the legacy server, hands it to a SAR (Segmentation and Reassembly) agent, and destroys itself. A distinct plugin object is therefore associated with each client interaction with the server.

Persistent application daemons can also be created at the proxy that either monitor legacy servers for updates, or receive *push*-style updates from the servers. The data from these updates is then handed to an application plugin, which hands over the data to SAR agents in the usual way. If the mobile is already connected to the proxy over a control channel, an application control message is sent to the mobile to notify it about pending data lying at the proxy. The mobile now either downloads the data into the *communications directory*, or instantiates the appropriate application plugin to handle the incoming data. If the mobile is not connected to the proxy, which could happen if the mobile is temporarily unable to access data services on cellular networks, an out-of-band SMS message can be sent to the mobile to indicate pending data. The client OCMP running on the mobile receives this SMS and tries to connect to the proxy whenever connection opportunities arise.

2.4.5 OCMP identifiers

As described earlier, OCMP identifies each mobile device by a unique GUID such as its IMSI [51]. The proxy uses this ID to demultiplex bundles belonging to different users. A different class of identifiers is needed for some applications. Consider a proxy that registers itself as the email server for a set of mobile users using a DNS MX record. When receiving incoming email, the proxy needs to find the user's OCMP-GUID. Therefore, the proxy needs to maintain a mapping from the user's application-specific address, such as an email address, to the user's GUID so that when the user connects to the proxy, it can send data to the correct user.

We have defined a framework on the proxy to support translation from application-IDs to OCMP-GUIDs. A registered application can create a daemon on the proxy that maintains mappings from application identifiers to the OCMP identifiers for all users of that application. This daemon is also registered to receive content from legacy servers. So, when a content server pushes data to the application daemon, it can instantiate an application specific plugin with the correct OCMP-GUID for the user, and redirect the incoming data to the plugin. The plugin caches the data in the usual way and delivers it to the mobile whenever it connects.

Note that each bundle carries a *GUID* to distinguish bundles belonging to different users, an *application identifier* so that bundles can be routed to the correct application, a *SAR agent identifier* for each data stream, and a *sequence number*. A concatenation of the first three identifiers defines a unique *session identifier*.

2.4.6 DTN support

DTN is modeled as a network interface for OCMP that implements its own transport layer protocol. Whenever a mobile device detects a WiFi network, OCMP examines the SSID to determine whether the network belongs to a DTN access-point, or it is a third party WiFi network providing a direct connection to the Internet (other methods can also be used, such a UDP broadcast on a particular port, or querying a GIS over the control channel for more information about the WiFi network). A different type of connection is instantiated depending upon whether the access-point provides a store-and-forward facility or not. We have implemented a new *Connection* object for DTN that talks to a DTN Bundle-Protocol-Agent (BPA) running on the DTN access-point. This is shown in Fig.2.2, and works as follows:

- *OCMP on mobile host*: A BPA does not run locally on the mobile host because the DTNRG reference implementation is not in Java. Instead, OCMP connects wirelessly to a BPA stub on the DTN access-point, and transfers data to it using TCP over the stub's RPC interface. Custody transfer acknowledgments are relayed back from the BPA to the DTN *Connection* object in OCMP. Thus, OCMP is made to believe that it is actually talking to a proxy in the Internet, but the BPA successfully masks the absence of an end-to-end route to the Internet.

- *DTN on access point*: The BPA receives bundles from OCMP, and stores them locally for subsequent forwarding to the OCMP proxy over DTN.

- *OCMP on proxy*: The proxy has a corresponding BPA running locally that receives DTN bundles from the access-points. OCMP bundles are extracted from the DTN bundles, and passed on to application plugins just like other OCMP bundles.

  For data to be sent to a mobile device through a DTN access-point, the scheduler on the proxy first checks whether the mobile device is registered with a DTN access-point, or it is likely to be in the presence of one. This step is crucial for the proxy to decide whether to retain the data in OCMP or to push it into the DTN overlay by routing it to an appropriate access-point. In the former case OCMP layers itself on TCP/IP as usual, but for the latter case, the proxy instantiates a connection to the BPA running locally and dispatches all the data to this BPA for eventual delivery to the user's custodian.

2.4.7 Control channel

Based on the discussions above, we summarize the uses of the control channel:

1. Send disconnection notifications when links terminate uncleanly, so as to reduce the amount of redundant state being maintained, and to exchange session state between the end points to reduce the size of the resequencing buffers.
2. Query the policy module on the mobile device for information about user preferences.
3. Exchange data required for policy enforcement algorithms. This data may include network performance status, remaining battery life on the mobile device, network cost information, etc.
4. Transfer application control data to instantiate plugins.

However, availability of the cellular network is not essential for correctness in OCMP. It is meant only as an optimization mechanism. We quantify some benefits of this approach in Section 2.6.

2.4.8 Policy control

The focus of this chapter is on an *architecture* that supports policy definitions. To exercise the architecture, we evaluate a simple policy described next (we are studying more sophisticated policies in current work).

We have implemented a simple policy enforcement algorithm for messages with deadlines, which works as follows. We connect to WiFi networks in preference to cellular networks, assuming lower usage costs for WiFi. We assume having

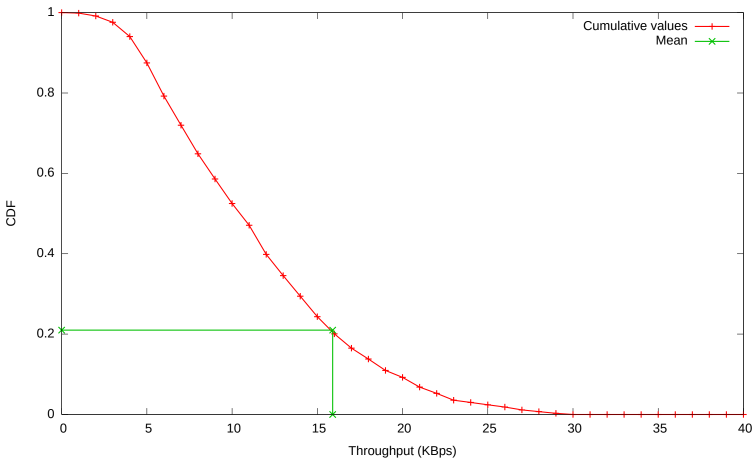


Figure 2.4: CCDF of EDGE application layer throughput

prior knowledge of the average throughput provided by the cellular network. This assumption is easily justified by an experiment, where an EDGE network was regularly probed over the duration of one day, and found to have a throughput of 16 KBps or more approximately 80% of the time, shown in Fig.2.4.

Each application data stream registers itself with the policy module on the mobile device, and specifies the size of its data stream, a delivery deadline, and direction of data transfer(uplink or downlink). The scheduler then back-computes an approximate commencement time for each data stream, ordering the streams from the furthest deadline to the nearest deadline. This is done by assuming the worst case scenario, i.e. the device may not run into any WiFi network and the cellular network will have to be used for data transfer; thus, data delivery must commence for each data stream preceding its deadline by a time of at least (size of remaining data / cellular throughput). If the commencement time for a data stream overlaps with the deadline for another data stream preceding it in the stream ordering, then the commencement time is appropriately adjusted to accommodate both the streams.

This is shown in Fig.2.5 as a *threshold envelope* for data delivery: the cellular network will be used only when the amount of data remaining to be transferred exceeds the envelope. The envelope can be computed at any instant of time using dynamic programming, given the deadlines and amount of remaining data for the applications currently in progress. For example, Fig.2.5 shows the commencement times for five applications ordered according to deadlines. The cellular network throughput is assumed to be 16KBps ($\sim 1MB$ per minute). The figure shows that since the commencement time computed for App-2 overlaps with the deadline for App-1, the corresponding threshold envelope is computed with an earlier commencement time. A similar adjustment is made for App-4 and App-5. In general, given n applications with corresponding deadlines and sizes of remaining data (d_i, s_i), ordered according to deadline such that $d_i \leq d_{i+1}$, and given the mean cellular throughput f, the commencement times (c_i) can be iteratively computed as:

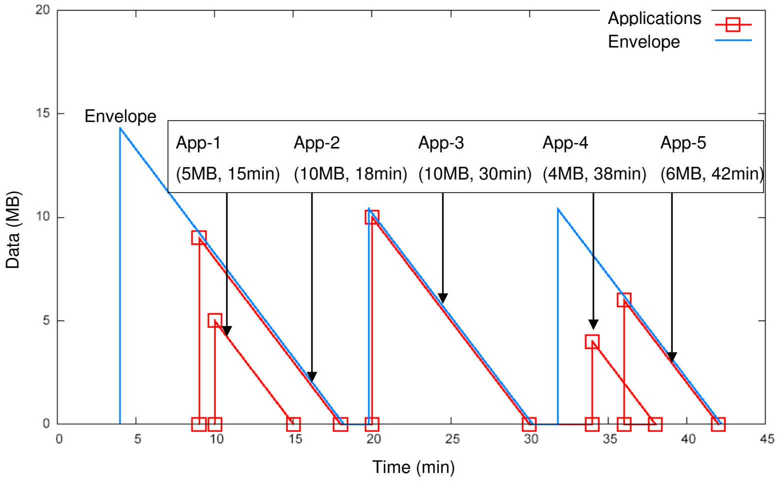


Figure 2.5: Simple network selection algorithm

if $\quad c_i \geq d_{i-1}$, then $c_i = d_i - f s_i$
else $\quad c_{i-1} = d_i - f(s_i + s_{i-1})$ and the process is repeated.

The threshold envelope is recomputed whenever a new data stream is registered, or a disconnection takes place from a WiFi network that was being used for data transfer. Timers are then initialized for each commencement time, so that data delivery can be triggered at that time instant. Now, whenever a new WiFi network is seen and there is pending data waiting for delivery, the WiFi network is always used for data transfer. On the other hand, the cellular network is used only when the remaining data exceeds the commencement time and there is no WiFi network available, or data is striped across both cellular and WiFi networks in case both are available. At any instant of time, the data stream with an earlier deadline is given preference over other data streams.

2.5 Implementation

We encountered a number of implementation problems in going from a paper design to a working prototype. These problems are described in detail in [185]. Here, we mention only one non-trivial roadblock.

Both the CLDC and CDC configurations for J2ME did not support the java.nio network library that provides a rich API for multi-interface communication. This led to problems in supporting simultaneous communication over multiple network interfaces to (a) enable data striping over cellular and WiFi interfaces, and (b) implement a control channel over the cellular interface, while transferring data on

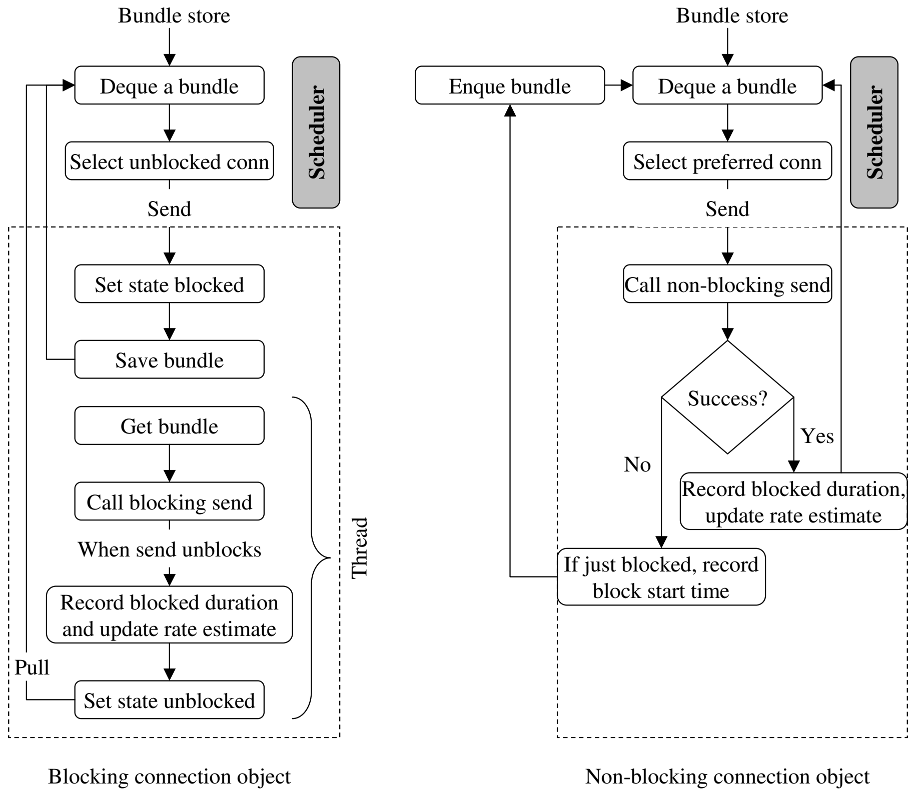


Figure 2.6: State maintenance with TCP blocking and non-blocking transport layer implementations

WiFi. Therefore, we implemented two thread-based variations through blocking and non-blocking write calls for TCP sockets, as shown in Fig. 2.6. The blocking version required a continuously looping thread in the scheduler plus one thread per interface, while the non-blocking version only required a single thread. Although the non-blocking version may appear to be better for this reason, but we found that the blocking version was able to better estimate the transmit rates on each interface, and hence resulted in smaller re-sequencing buffer sizes at the receiver when striping data across multiple interfaces. Good estimation of the transmit rates was necessary because the cellular and WiFi channels were widely dissimilar in delay and throughput, and hence the bundle transmission sequence had to be matched to maintain approximately in-order delivery at the receiver. A second reason to maintain these statistics was to make them available for policy decisions on network selection. The rate estimation was done through the standard EWMA technique using a forgetting factor of 0.8, and resulted in data-striping scenarios yielding the optimal aggregate performance of individual non-striping scenarios on either interface.

2.6 Evaluation

We used an HP-Compaq 1.8GHz laptop running Windows XP for our experiments. A WiFi connection was provided over 802.11g with a DSL backhaul of 5Mbps, and a cellular connection was provided through an EDGE PCMCIA card. All our experiments were conducted in a stationary environment to minimize the effects of mobility on our results. We instead implemented an emulation-module in OCMP to emulate WiFi connections and disconnections, assuming an arbitrary mobility schedule. We recognize the fact that our emulation methodology does not take into account factors such as 802.11 association delays, wireless losses due to interference and mobility, automatic rate adjustment, impact of lower layers on TCP and other transport layer implementations, etc. We defer an analysis of such factors to future work, and retain the focus of the evaluations in this chapter on the design of the policy based system that we have proposed.

2.6.1 Meeting the design goals

We evaluate how we have met our design goals stated in Section 2.2.

1. *User-directed use of multiple networks*:

   Fig.2.7 and Fig.2.8 show a runtime trace of the number of bytes that remain to be transferred for a single application during one emulation run, while using the simple algorithm of Section 2.4.8. In both the figures, grey regions denote the times when WiFi is absent. We have divided the timeline into multiple zones for clarity of explanation. We will use $r(t)$ to denote the number of

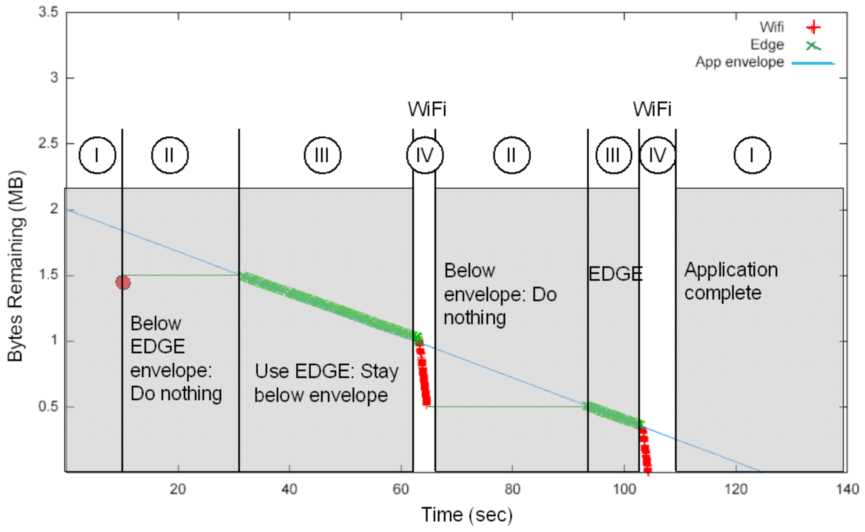


Figure 2.7: Single application, easy deadline

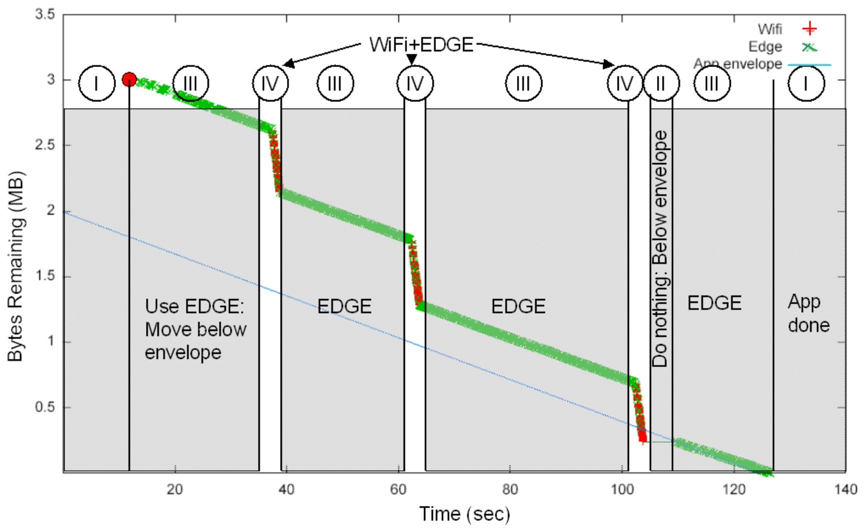


Figure 2.8: Single application, aggressive deadline

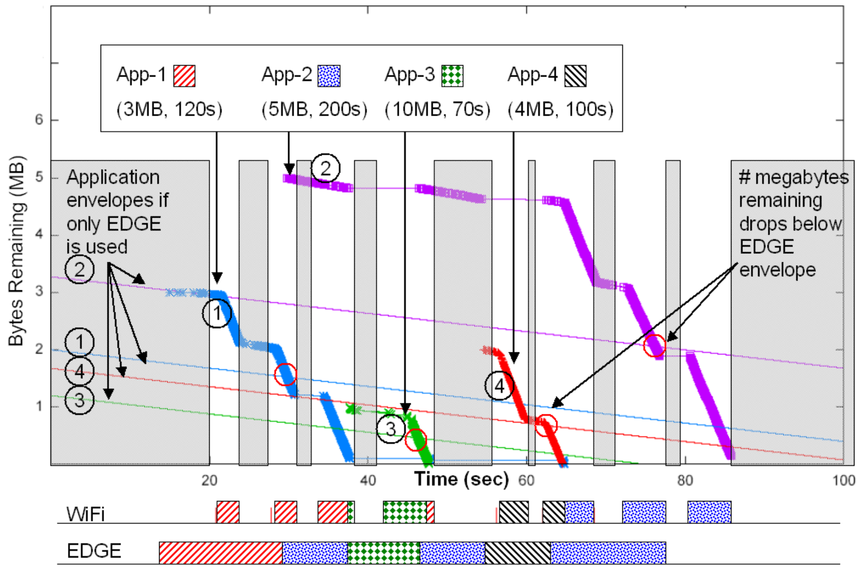


Figure 2.9: Multiple applications

bytes that remain to be transferred at time t. We call the scenario depicted in Fig.2.7 as having an easy deadline because when the application starts, the size of its data to be transferred (1.5 MB) is less than the maximum amount that can be sent on EDGE even if no WiFi network shows up. The scenario in Fig.2.8 is correspondingly termed as having an aggressive deadline because the amount of data to be transferred (3 MB) is larger than the maximum amount that can be sent on EDGE alone.

(a) *Zone I (Idle):* This denotes an idle state, either when there is no application waiting for data delivery, or all the applications have completed their respective data transfers.

(b) *Zone II (Do-nothing)*: This denotes the time during which an application is active, but there is no data transfer is progress because $r(t)$ is below the envelope. Thus, Zone II occurs in Fig.2.7 when the application is started at a time that is much before the commencement time calculated assuming that no WiFi network will be available until the application deadline.

(c) *Zone III (EDGE)*: This denotes the intervals during which there is no WiFi network available, and $r(t)$ will either cross the envelope if EDGE is not used, or is already above the envelope.

(d) *Zone IV (WiFi)*: This denotes the short intervals of time when WiFi networks are available for opportunistic use. In Fig.2.8, these opportunities present themselves when $r(t)$ is above the envelope; hence, both WiFi and EDGE are simultaneously used. However, only WiFi networks are used in Fig.2.7 because $r(t)$ always stays at or below the envelope.

Fig.2.9 shows a runtime trace for 4 applications running under the same policy enforcement algorithm, each application having a single data stream. All applications have been shown to have aggressive deadlines, where they start above the EDGE envelope that is available for them. Note that the applications are labeled according to their start times, but their deadlines are not in the same order: App-3 < App-4 < App-1 < App-2. At any time instant, the application having an earlier deadline is given preference over other applications. Thus, App-1 is allowed to use WiFi in preference to App-2 always, and App-3 preempts App-1 at $t = 38sec$ when App-3 starts and has an earlier deadline than App-1. The earliest-deadline-first ordering is also followed on EDGE. Thus, App-2 which has the last deadline, is able to use EDGE only when all other applications have either completed, or have dropped below their individual EDGE envelopes.

These experiments show that user and application-directed use of multiple interfaces is possible through OCMP.

2. *Ease of application design and implementation*: We believe we have successfully met this goal because both our plugin and directory APIs completely mask the effects of network disconnections and switching. As anecdotal evidence, a student without any knowledge of the underlying details of OCMP was able to develop an Email plugin in just 4 days, "simply by parsing /var/spool/mail and dumping the emails in the OCMP communication directory".

3. *Support for legacy servers*: We have so far built OCMP application plugins for the following services that are otherwise available only through applications that connect directly to the legacy servers hosting these services: send and receive email, upload photos to www.flickr.com, post blogs to www.blogger.com, and receive HTML pages using HTTP GET.

4. *Application session persistence across disconnections*: Instances such as the occurrence of Zone II in Fig.2.7 and Fig.2.8, when no networks are used, show that session persistence is supported in OCMP.

5. *Optimized network switching*: We stated earlier that in the absence of quick notifications of WiFi disconnections sent over the cellular control channel, the proxy will have to maintain a large amount of state until when a TCP timeout occurs. To observe this effect, we ran a set of 10 experimental runs both with and without an EDGE cellular channel, and calculated the mean and standard deviation of the delay incurred by the proxy in inferring a disconnection. We did this by manually disconnecting the laptop from WiFi and observed the delay until when the proxy closed the TCP connection. The results are shown in Table 2.

   When EDGE is used as a control channel, the total delays are of the order of 1 sec. The primary component of this delay is the large RTT of an EDGE

Table 2.2: Disconnection detection latencies

	Mean	Std. deviation
With EDGE	982ms	156ms
Without EDGE	49.2sec	13.7sec

network ($\sim 700ms$). Note that in these measurements, we did not count the latency incurred by the 802.11 MAC layer to infer a disconnection and report it to OCMP. This is because such link layer indications are not yet a part of the 802.11 standard [28], although they are useful for opportunistic communication. For now, we assume that such link layer triggers exist, and disconnections can be detected by the mobile device within a couple of milliseconds. This is because most wireless cards consider three consecutive MAC retransmission failures as a disconnection [52], which can be done within a few milliseconds.

On the other hand, without an EDGE control channel, TCP timeouts often take over a minute to detect a broken connection. This means, for example, that the proxy maintains flow state over a period of minutes even though opportunistic WiFi network residence times will likely to be of the order of a few tens of seconds [55]. This observation is of significant concern because the OCMP proxy will likely be shared among hundreds of users, and redundant state maintenance can clearly decrease the scalability of the proxy. This also shows that link layer indications can prove helpful for opportunistic communication, and argues for the inclusion of link layer triggers as a part of the 802.11 standard.

6. *Buffered Access Point support*: We have extensively tested the integration of OCMP and DTN with encouraging results obtained on Soekris boxes running as access points. We also plan to repeat our experiments with off-the-shelf AP hardware.

2.7 Discussion and related work

We are not aware of any other work that provides the same set of functionality provided by our system. However, our work is closely related to, and builds on the insights of several threads of past work in this area, as described next.

Policy-based selection of network interfaces was first introduced in [44], and has since been extensively explored in the context of vertical handoffs by researchers [45, 46]. The problem was motivated by the desire to choose the best network that optimizes metrics such as data rates or power consumption in the long term, while preserving seamless connectivity. We have built upon this definition by noticing that seamless connectivity is not required for many non-interactive applications. Therefore, our policy definitions operate at the session layer, unlike the network-

and transport-layer policies of past work. This allows us to include a delay component (such as a deadline for data transfer) at a timescale of minutes and hours, which fundamentally alters the system.

Our use of many wireless interfaces in parallel is similar in spirit to pTCP [36]. Unlike pTCP, which assumes an underlying TCP connection, OCMP is transport-agnostic. We are able to make this tradeoff because we are primarily interested in delay-tolerant applications that can deal with a large resequencing buffer. In contrast, pTCP supports interactive applications, and must therefore exploit TCP structure to reduce the size of the resequencing buffer. Unlike pTCP, we have built, deployed and evaluated the performance of the system in a testbed, instead of relying on simulations.

The use of location-independent identifiers for resuming a session was proposed earlier in the context of Rocks-and-Racks [53] and TCP Migrate [42]. However, these earlier solutions are not only TCP-centric but also support only a single interface. Our use of an almost-always-available cellular connection for the transmission of control messages (i.e. data available, and link down) distinguishes us from these proposals. Also, unlike these proposals, we have designed and implemented a session-level reliability protocol.

Our use of a proxy for dealing with session disconnections and the aggregation of multiple transport connections into a single connection is similar to that proposed in PCMP [41]. However, OCMP differs from PCMP in several ways. First, unlike PCMP, OCMP supports the use of multiple NICs in parallel. We also support arbitrary transport protocols including UDP with erasure codes [29, 30], and TCP optimized for cellular networks [31, 39], whereas PCMP is essentially TCP-centric. Unlike PCMP, OCMP nodes can be powered down because application data and control is persistently stored. Our architecture allows session state to be encapsulated and transferred from one proxy to another. This allows us to reassign a mobile to the closest available proxy, greatly improving performance. Finally, servers can push data to OCMP proxies, or plugin daemons can poll legacy servers to pull data, and the data can then be picked up opportunistically by mobile devices. We believe that these differences make OCMP more suited to a dynamic multi-network environment than PCMP.

Our proposal to use multiple OCMP proxies for better throughput during opportunistic connections is similar to that proposed in DHARMA [26]. DHARMA uses a network of distributed *home-agents* to provide (a) mobile IP like packet forwarding support when client IP addresses change due to mobility, and (b) session persistence for preserving TCP connections across disconnections. We believe that our work is more general because it provides many more features than just session persistence and mobility support.

Intelligent selection of network interfaces with session persistence is also being explored in the context of the Haggle project [37]. However, Haggle is focused on infrastructure-less systems where devices communicate with each other in an ad-hoc manner. Further, they do have a notion of application plugins, proxies, or a control channel.

Finally, our work is complementary to, and extends, recent work in the area of implementing a router for delay tolerant networks (DTN) [32]. Our notions of session persistence, data persistence, bundling, and multi-network support originate in this seminal work. However, we have made several non-trivial extensions. These include the support for fine-grained policy control, the notion of application plugins, the use of a proxy, and the separation of the data and control planes. Some of our detailed design decisions also differ from that made in the DTN reference implementation. For instance, DTN associates a *convergence layer* with each transport protocol, which means that all NICs that support TCP would use the same convergence layer. In contrast, we associate a connection with each NIC, allowing us to exploit network heterogeneity by optimizing different transport layer implementations for different types of networks [29–31,39]. OCMP also supports a control channel to communicate disconnection and reliable data transfer information between peers, which is not currently possible in DTN. We observe that our work is motivated by a narrower set of problem areas than DTN, which allows us to exploit the inherent problem structure to make these optimizations.

There are many open areas of research related to opportunistic communication. How can a mobile device detect networks in a power efficient manner, without having to keep its radio switched on all the time? What is the interaction between TCP and lower layers during an opportunistic connection interval? What kind of transport layer implementations can efficiently handle these brief connection opportunities? How should wireless devices be designed to make opportunistic communication more power efficient? We plan to address these issues in future work.

Chapter 3

The KioskNet project

The less educated you are, the more bandwidth you need to communicate –
APJ Abdul Kalam, 2006

In the previous chapter, we showed how opportunistic connectivity through mobile devices could significantly reduce the communication cost for delay tolerant applications. In this chapter, we use the same principles to develop a system for enabling low-cost connectivity in rural areas.

A popular mechanism to provide rural connectivity in many developing countries is through public access terminals in shared **kiosks** in villages. An Internet connection is provided to a public kiosk in a village, rather than to each individual in the village. This shared infrastructure automatically amortizes the cost of connectivity for the kiosk across all the people who use the kiosk. Today, kiosks connect primarily using dialup lines, satellite Very Small Aperture Terminals (VSAT), or rarely, long-range WiFi. Unfortunately all three solutions suffer from insuperable problems. Dialup land lines are slow and unreliable. In rural areas, repair delays of three to four days are common. VSAT terminals are reliable, but require considerable up-front capital expenditure as well as expensive monthly fees. Their cost-per-bit is therefore high. Moreover, in case of breakdowns, spare parts are not widely available. Finally, long-range WiFi requires considerable planning for a large scale deployment. Line-of-sight is necessary for most technologies, and a constant power supply is needed at each relay tower. Early adopters such as N-Logue and Drishtee in India report unexpected problems such as a long-range link being overpowered by a laptop near one of the towers. In practice, tower heights of at least 18m have been reported to be necessary, which turn out to be quite expensive [63]. Looking to the future, it is likely that high-speed 3G cell phone technologies such as EvDO will eventually reach rural areas. However, rural areas are usually poor, so it is unlikely that high-speed data services from cellular providers will be offered any time soon.

Given this situation, our interest is in providing low-cost and reliable connectivity to rural kiosks. In this chapter, we present an architecture that uses buses and cars (*mechanical backhaul*) to ferry data to and from a kiosk, building on the pioneering work of Daknet [66]. This design decision has repercussions on every layer of the protocol stack, and indeed, the entire network architecture. We therefore present principles for naming, addressing, forwarding, and routing that are needed by any system that uses mechanical backhaul. We also report on our experiences with implementing this architecture in the context of the Delay Tolerant Networking architecture proposed by the IRTF DTN Research Group [71]. We even deployed a first prototype in the field in May 2006.

The chapter is laid out as follows. We present our design goals in Section 3.1 and outline why we cannot use existing research to address this problem. We then describe in Section 3.2 the technologies available to reach the goals, and enumerate fundamental principles on which our work is based. We give an overview of our architecture in Section 3.3, and describe details of location management, routing, and protocol design in Section 3.4. Details from a pilot deployment in India are presented in Section 3.5. Finally, we describe related work in Section 3.6.

3.1 Design goals

We first present the goals of our work.

- **Low cost**: To be widely deployed, our system has to be low-cost. We would like to keep the capital cost per kiosk to be under US $250 and the operational costs to be as low as possible. For example, we would like a kiosk supporting a user base of about 500 users to charge no more than US$0.15/month for email service.

- **Reliable**: To be useful, the system has to be reliable, tolerating power outages at kiosks, ferry failures, packet loss, and disk corruption. Moreover, this reliability has to be designed in, because the system will be operated by technically untrained users.

- **High bandwidth**: The system has to be scalable to support applications that require large amounts of data to be transferred to and from the Internet. For example, digital photographs uploaded from a kiosk can be more than 2 MB in size, and cannot be easily uploaded over a dialup connection. Video clips are even larger.

- **Disconnection tolerant**: We require communications to be disconnection tolerant for two reasons. First, this allows the system to work reliably despite power outages, which are endemic in developing countries. Second, it allows us to trade delay for cost. That is, we can reduce costs by tolerating message delays of up to a few days. Because of disconnections, both ends of a transport connection may not be simultaneously present [35], precluding the use of standard TCP/IP to provide end-to-end connectivity.

- **Allow user mobility**: Field studies show that many villagers routinely move from village to village within a 15-20km radius [63]. We would like to allow such villagers (or health care workers, who also move from village to village) to access their data from the closest kiosk.

- **Interoperability**: Clearly, applications such as VoIP are incompatible with mechanical backhaul. Nevertheless, to the degree possible, we would like the users to be able to access Internet services on legacy servers *without modification to these servers.*

- **Policy based use of all available networks**: We would like a kiosk to be able to use all channels of communication (including cell phone and dialup) based upon intelligent policies. For example, some applications might require immediate data transfer and dialup may be the only option possible, as opposed to email or land-record applications that are inherently tolerant to delays and more suited to use a mechanical backhaul communication mechanism.

- **Support both kiosk and laptop/pda users**: We envision that some users will be using shared PCs in a kiosk to access networked services, while other users may have their own device; the latter users may be farmers or NGO employees who may have a PDA or another mobile device, or even a cellular smartphone that is out of coverage area in villages. We would like to support both classes of users.

Do we need a new solution?

Before describing our solution, it is worth considering if the goals presented here can be achieved using existing research. First off, it is clear that a legacy solution, that is, naming nodes with DNS names, addressing them with IP addresses, and using TCP end-to-end, will not work, because the two-ends of a connection, i.e. kiosk users and legacy servers, are never simultaneously present [35]. Nevertheless, recent research presents solutions which on surface appear to meet all our goals. For instance, past work has addressed disconnection tolerance [42], support for mobile users [34, 40, 67], interoperability with legacy servers using proxies [65], and use of multiple networks and NICs [61].

However, all these solutions have three problems. First, they are point solutions that do not form a single coherent system. It is not possible to simply mix and match the solutions to achieve our goals. Second, they have been designed in the context of laptops and PDAs that are almost always connected, with short disconnected periods, and, when connected, are connected at relatively high speeds. Finally, they do not focus on low cost and reliability. Introducing these design constraints greatly changes the form of the solution.

This motivates us to seek a custom-built solution to our problem, using, where possible, design principles advocated in existing research. We return to a more detailed evaluation of past work in Section 3.6.

3.2 Available technologies

We leverage the following fundamental technologies to meet our goals:

- **Cheap storage**: Storage today is cheap, costing less than US$1/GB, and rapidly getting cheaper. Therefore, we envision several tens of GB of storage at a kiosk, on a bus, and anywhere else needed in the network to store data in transit.
- **Wireless networks**: Wireless (802.11x) network cards are ubiquitous, cheap, and, for the most part reliable. Wireless allows us to make opportunistic use of buses and cars that happen to drive past a kiosk, and then exchange data as they drive past a server that has a persistent Internet connection.
- **Delay Tolerant Networking overlay**: The DTN Research Group architecture provides a delay- and disruption-tolerant bundle-forwarding architecture. At its core, the architecture describes how a set of DTN routers form an overlay to cooperatively store and forward *bundles* of information [35]. DTN routers are connected by links that are sometimes, or often, down, and can be persistent, scheduled, or opportunistic. The DTN architecture is ideally suited to our needs because it supports the opportunistic and scheduled links provided by buses. Although DTN routing schemes are yet to be defined, its naming and addressing conventions are simple and extensible, and the bundle forwarding engines are available as open source. We have therefore built our design as an extension to this architecture.
- **Cellular overlay**: Unlike 3G data services, which are expensive to deploy, GPRS-like data services at low bit rates (4-8 kbps) are nearly ubiquitous worldwide, even in rural areas. It appears to be straightforward to use recycled cell phones as GPRS modems. We therefore seek to exploit this connectivity, where available, to provide a cheap and reliable control plane.

Design principles

Before embarking on any architectural design, it is useful to identify the principles embodied in the architecture. These principles allow us to intelligently navigate the infinite space of possible designs. We believe these principles are applicable to any realistic architecture that uses mechanical backhaul and has goals substantially similar to ours. We hasten to add that we do not claim these principles to be novel; instead, we claim that these are the principles *relevant* to our context.

- **Lowered cost through shared infrastructure**: Low cost can only be achieved by sharing every component of the architecture. Therefore, we need to share not only the Internet gateway, but also the storage on the bus, and every element in the kiosk. Of course, unrestricted sharing can be unfair. Therefore, all shared elements need to be appropriately managed.

  Note that the proposed One Laptop Per Child project [72] does not share end-systems. Therefore, we do not believe that this project can achieve the cost targets achievable using shared kiosks.

- **Store and forward of self-describing data**: This is necessary for tolerating disconnections and disruptions. Store and forward allows a node to deal with failed links. Moreover, making bundles self-describing, in the same way that an email message is self-describing, allows easy recovery from power failures at nodes and from bad routing decisions.

- **Decoupling location and addressing**: Because users are mobile, their identifiers must be location-independent. This means that we need some way to map from a user's ID (which is the eventual destination of a data packet) to the ID of his or her current location (i.e. address).

- **Opportunistic link use**: The opportunistic use of links increases sharing of infrastructure nodes such as kiosks and buses and thus reduces cost. Moreover, this principle dictates that we should attempt to use every available NIC at a kiosk based upon intelligent policies. Therefore, kiosks should exploit not only buses, but also dialup links, and VSAT connections, whenever required.

- **Separate data and control plane**: The clean separation of the data and control planes allows us to use almost-always available cellphone links for the control plane, and opportunistic WiFi or Bluetooth links for the data plane. By using costly cell phone links for low-bandwidth communication, we optimize the usage of the data plane. In particular, using the lightweight control plane for routing updates allows us to overcome pathological routing problems that arise when routing or location updates are delayed.

- **Proxies for legacy support**: We use disconnection-aware proxies to allow applications on a kiosk to interoperate with existing Internet applications.

- **Replication for reliability**: We can easily replicate data on disks for reliability. We even contemplate replicating bundles on the network. Disk is cheap enough that this redundancy is unremarkable. However, it is not clear whether, and to what degree, packet replication on the network is needed. Finally, we envisage that ferries can carry a set of spare parts that can be used, as needed, by kiosks along its path.

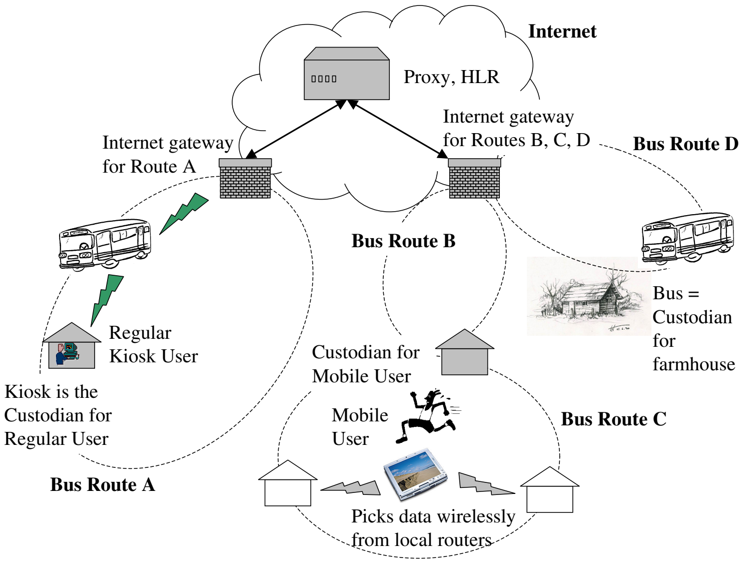

Internet
Proxy, HLR
Internet gateway for Route A
Internet gateway for Routes B, C, D
Bus Route D
Bus Route B
Bus = Custodian for farmhouse
Regular Kiosk User
Custodian for Mobile User
Kiosk is the Custodian for Regular User
Mobile User
Bus Route C
Bus Route A
Picks data wirelessly from local routers

Figure 3.1: Example topologies

3.3 Design overview

This section presents an overview of our architecture outlining the data and control flow in our system. Details are presented in subsequent sections.

Kiosks play a central role in our architecture. A kiosk consists of a *kiosk controller*, a server that provides network boot, a network file system, user management, and network connectivity by means of dialup, VSAT, or mechanical backhaul. A kiosk controller is assumed to have a WiFi NIC, and very likely a GPRS/EDGE or dialup connection. Our prototype uses headless and keyboard-less low-power single-board computers from Soekris Corp. as kiosk controllers, although the functionality can be implemented in any commodity PC. Our choice of a Soekris device was due to that fact that it only draws 7W, and can therefore be powered by a battery-backed solar cell. Moreover, the lack of I/O devices discourages tampering.

The kiosk is expected to be used by two types of users, shown in Fig.3.1. Most users would use a public access terminal that boots over the network (using RAM disk) from the kiosk controller, and can then access and execute application binaries provided by the kiosk controller over NFS. Recycled PCs that cost around $100 are ideally suited to function as public access terminals, and spare parts are widely available worldwide too. Moreover, as a shared resource, they are an order of magnitude cheaper than any dedicated resource.

A second class of users, such as wealthier villagers, government officials, or NGO partners, could access one or more kiosks, or a bus directly, using their own devices, such as smart phones, PDAs, and other mobile devices. Such users would use the kiosk-controller or bus essentially as a wireless hotspot that provides store-and-forward access to the Internet. Identity management and mobility support for hotspot users is a key requirement supported by our architecture.

Although kiosk controllers can communicate with the Internet using a variety of connectivity options, our focus is on the use of mechanical backhaul. This is provided by cars, buses, motorcycles, trains, and even bullock carts that pass by a kiosk and also an internet gateway (which is described in more detail below). We call such entities *ferries*. A ferry has a small, rechargeable, battery-powered computer with 20-40GB of storage and a WiFi card. It opportunistically communicates with the kiosk controllers and internet gateways on its path. During an opportunistic communication session, which may last from 20 seconds to 5 minutes, anywhere from 100KB-50MB of data can be transferred in each direction. This data is stored and forwarded in the form of self-identifying *bundles*.

Ferries upload and download data opportunistically to and from an *Internet gateway*, which is a computer that has a WiFi interface, storage, and an always-on connection to the Internet. The gateways are likely to be present in cities having DSL or cable broadband Internet access. A gateway collects data opportunistically from a ferry and stages it in local storage before uploading it to the Internet. It also downloads bundles on behalf of kiosk users, and transfers them opportunistically to the appropriate ferry, governed by a routing protocol. In our implementation, we use a Soekris device both for ferries and for gateways.

We use the term *DTN router* to refer to any device that is connected to more than one other device either on different NICs, or at different points in time. In this sense, the ferry is just a mobile DTN router, and Internet gateways and kiosk controllers are examples of fixed DTN routers.

We expect that most communication between a kiosk user and the Internet would be to use services such as Email, audio and video downloads, financial transactions, and access to back end systems that provide government-to-citizen services, such as land record management, birth certificates, and various sorts of licenses. Systems that provide such services are typically unable to deal well with delays and disconnections. Therefore, we propose the use of a disconnection-aware proxy that hides disconnection from legacy servers. The proxy is resident in the Internet and essentially has two halves. One half establishes disconnection-tolerant connection sessions with applications running on a recycled PC or on mobile user's device. The other half communicates with legacy servers. Data forwarding between the two halves is highly application dependent. For example, a proxy that fetches email from a POP server on behalf of a user needs to implement the POP protocol. To support application-specific protocols, we allow applications to instantiate an application-specific plugin at the proxy. Our system can support multiple proxies; for each gateway we use the proxy closest to it in terms of the RTT between them so as to gain maximum TCP throughput. In the rest of the discussion we assume a single proxy to keep the explanations simple.

Finally, the last component of our architecture are the legacy servers that are

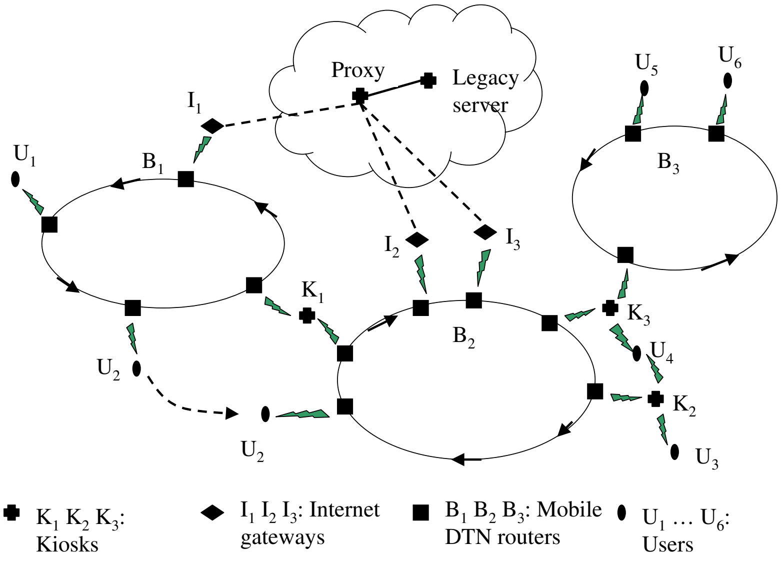


Figure 3.2: Network model

typically accessed using TCP/IP and an application-layer protocol such as POP, SMTP, or HTTP by a proxy. We do not require any changes to legacy servers.

3.3.1 Network model

We model the system as shown in Fig.3.2. The Internet IP core is a fully connected cloud where all nodes form an overlay clique. The core is connected using low-speed links to the Internet gateways I_1-I_3. Each ferry on a particular route is represented by a ring of nodes, with each node representing a point in the ferry's trajectory where it communicates with a kiosk or user. The weight of an edge on the ring can be used to represent the transit time between contacts. Note that some ferry routes go past multiple Internet gateways, while others go past none.

Kiosks such as K_1 - K_3 hang off the ferry ring. Some kiosks, such as K_1 and K_3 can be used to route bundles from one bus route to another. Finally, users are attached to a single kiosk (e.g. U_3), to multiple kiosks (e.g. U_4), directly to a ferry (e.g.. U_1), or directly to multiple ferries (e.g. U_2). For the purpose of routing, we distinguish between users who always access a ferry at a particular point in its trajectory, such as from a farm house, and users who opportunistically download data from a ferry at some (potentially varying) point in its trajectory. We represent the former as a node along the ferry's trajectory because it is possible to speak meaningfully about paths and edge weights. The latter are not represented in the graph, and, for routing, we treat them as if they are located on the ferry itself. If such a user were to move from one ferry to another, we update their location from one ferry to another.

The performance achievable by this system can be characterized by two metrics:

the overall client-to-server throughput achievable, and the mean end-to-end delay. In terms of throughput, the path from a legacy server or proxy to an Internet gateway is highly constrained by the speed of the gateway's access link [66]. This link operates typically at 100 kbps over a DSL connection. Therefore, to maximize performance, we should keep the Internet gateways as busy as possible, balancing load amongst all available gateways.

In terms of end-to-end delay, naturally, ferry transfer latencies can add hours or days to a communication path. Surprisingly, a significant contributor to end-to-end delay is not only the ferry transit time, but also the wait time at an Internet gateway awaiting upload. To see this, note that a bus can pick up 20MB at each kiosk it visits [66], so, if it visits 10 kiosks, which we expect to be the median, it would pick up 200 MB. Of this, perhaps 80% or 160MB would need to be transferred over an Internet link. At 100 kbps, this will take nearly 4 hours. If more than one bus were to share a single gateway, the delay can be substantially larger. So, with a poor choice of routing, a gateway could introduce a few days worth of delay for data arriving on a single ferry contact!

3.3.2 Protocol architecture and data path

We now trace the flow of data from a client to a legacy server, shown in Fig.3.3. The client software application executes either on a mobile device or on a recycled PC, and the subsequent discussion applies equally to both situations.

Applications may either be aware of our architecture, or not. If they are, then the application directly communicates with the Opportunistic Communication Management Protocol (OCMP) daemon, described in the previous chapter. OCMP maintains a client-to-server disconnection-tolerant session with the help of application-specific plugins running at the client and at the proxy. Besides maintaining a disconnection-tolerant session, OCMP also manages multiple client NICs and provides application-specific ID to end-system ID translation at the proxy.

In order to support disconnection-tolerant applications on the client that are not aware of OCMP, we run a modified dummy server on the mobile itself, the kiosk-controller, or the ferry. This server pretends to be the legacy server. For instance, this server could be a Jabber server or an email server. On receiving data from the client, the modified server invokes OCMP which encapsulates client data into bundles. This allows us to support legacy client applications with no modifications. Of course, the application needs to be inherently disconnection-tolerant – our approach cannot mask the inherent delays in the communication path.

Once OCMP receives data, it stores it in a local on-disk database. This allows it to gracefully tolerate power disruptions that may bring down the kiosk controller, perhaps as often as several times a day.

When an opportunistic connection presents itself, OCMP hands bundles for transmission to a DTN Bundle Protocol Agent [60], which invokes a routing and

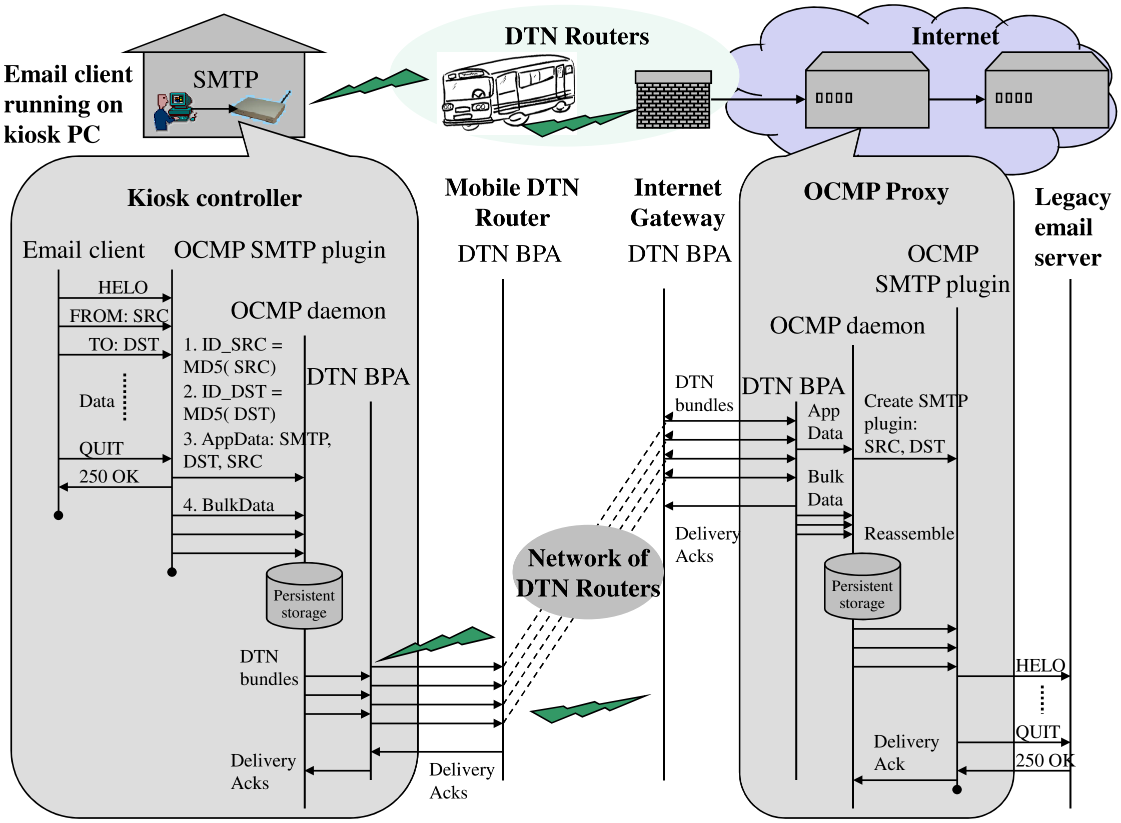


Figure 3.3: Data path

flow control protocol (described in Section 3.4.4) to decide which bundles to transfer on the opportunistic link. The selected bundles are transported, typically using TCP/IP, to a DTN router on the ferry. We use the standard bundle protocol for this transfer, as described in [60]. When the ferry goes past an Internet gateway, or a kiosk is used to route between ferries, routing and flow control are again invoked to select bundles for transmission to the gateway. These are then transferred to the gateway using the bundle protocol. The destination of the bundles is either one of the proxies (for legacy applications) or the bundle-aware server. The Internet gateway accepts custody of the bundles and schedules them for transmission on the internet link, keeping in mind bundle priority and any other scheduling decisions.

Bundles meant for legacy servers that arrive at a proxy are demultiplexed and handed to the appropriate application-specific plugin. This plugin then invokes a connection to the legacy server over TCP/IP and delivers the data.

Data flowing in the reverse direction is symmetric. The proxy registers itself on behalf of the clients to receive, for example, email destined to them. On receiving data, the application-plugin on the proxy establishes a disconnection-tolerant connection to the client and queues bundles for transmission. These are then delivered to the OCMP stack running either on the client or the kiosk-controller using OCMP layered over DTN, and thence to the client application.

The next few sections expand on this overview. We start with a detailed description of naming, addressing, and location management.

3.4 System architecture

3.4.1 Naming, addressing, and location management

In our system, a user's human-readable name is his or her telephone number (IMSI) or email address. For uniformity, the system translates both into 20-byte strings with a SHA-1 hash. We call such a string a Globally Unique ID or GUID. Because of its length, we assume that users with distinct names will almost surely map to distinct GUIDs. We also use the same GUIDs for forwarding, thus our system forwards bundles using names rather than addresses.

Creating GUIDs from a hash of a human readable name allows translation from a human-readable name to a fixed-length numeric GUID to be carried out without any additional infrastructure. In contrast, with DNS names, a sender needs to use the DNS service to find its correspondent's IP address, and with HIP, a sender needs to determine the cryptographically signed ID of its correspondent using PKI. This choice of GUIDs is also motivated by our security architecture described in the next chapter, where the SHA-1 hash of a user's name is also that user's public key.

Every user in our system registers itself with at least one DTN router at a kiosk, bus, or Internet gateway, called its *custodian*. A custodian is similar to a mail server in an email system, in that it holds data on behalf of a potentially disconnected user, and that it participates in a routing protocol to forward bundles between users. A custodian can be thought of as a rendezvous point that coordinates a sender and receiver, in the same way as a Foreign Agent in Mobile IP [67], an anchor point in Hierarchical Mobile IP [69], or an I3 server in I3 [34]. Note that from the perspective of the system, as long as bundles are delivered to the custodian, they will be somehow picked up by the user. Custodians play multiple roles in our architecture. They store bundles on behalf of a user to be picked up later. They also act as an anchor point for mobile users and as a gateway to limit the spread of routing updates. A user's custodian must be able to reach that user either directly, i.e. on a one-hop link, or by means of a series of DTN routers that need only the user's GUID to deliver bundles to it. Each custodian keeps track of the user GUIDs that are registered with it, and whether the bundles will be directly picked up by the user or they have to be routed further. If the bundles have to be routed further, then a bit is flipped in the bundle header to indicate that the bundles are to be forwarded to the user rather than to the intermediate custodian.

The full *address* of a user is the tuple with two names: [custodian GUID, user GUID]. If a sender does not know the custodian GUID, it can use the special form [*unbound*, user GUID]. This instructs a DTN router to forward the bundle on its default path to a router that can eventually find the user's custodian and forward bundles to it. To resolve unbound bundles, we assume that a redundant and fault tolerant server on the Internet (which may be implemented as a multi-site cluster or a distributed hash table for scalability) stores a lookup table or registry with a mapping from a user's GUID to its custodian's GUID. Following cellular telephony terminology, we call this lookup table the Home Location Register (HLR). A user can simultaneously register with multiple custodians; in this case, the HLR resolves the user's GUID to the GUIDs of multiple custodians, all of which represent viable routes. A user can also be identified by multiple GUIDs based on different public identifiers. In this case, the HLR maintains records mapping the different GUIDs.

When a user moves, the HLR is updated if and only if the user changes her custodian. Users can potentially send location updates anticipating future movements if they know their eventual destination, so that the data can be forwarded in advance to the closest custodian.

Fig.3.1 shows three example cases. In the first case, a static user who visits a single kiosk defines the associated kiosk controller as her custodian and registers with it. The controller then retains all data for the user locally. In the second case, a user who moves between neighboring villages define her custodian as a DTN router higher up in the hierarchy towards the Internet. In this case, the custodian uses a routing protocol (described in Section 3.4.2) to direct bundles to the user. In the third case, a mobile DTN router, such as a bus, is defined as the custodian, for example, to deliver data to a rural farmhouse PC that is far from any village kiosk.

Setting up the HLR

The HLR keeps track of the custodian(s) associated with each user. How should it be kept up to date by a potentially disconnected user? We first describe a solution that assumes that every user is associated with a single custodian.

In case of a user who will always access the system from a single kiosk (the common case), when a user is added to the system, the user's ID is registered with the kiosk controller, which is also its local DTN router. In the case of a mobile user, it opportunistically associates with any available DTN router, which we call its local DTN router.

On association, the local DTN router updates the routing protocol (for instance by creating a Link State Packet) to indicate that it can reach the user. The router also informs the user of its choice of nearby custodians (which may be a nearby kiosk, a custodian in the Internet, or the bus itself). The user chooses one or more custodians and informs the local DTN router of its choice.

At a future time, the DTN router informs the custodian of the user's request. The custodian updates its state to indicate that the user is now registered with it. If the custodian has not seen the user before, then the HLR needs to be updated to reflect the new custodian choice. The custodian updates the HLR by sending a registration message to the HLR. This message is also sent to the user's previous custodian to free up any old bundles: these old bundles are then sent to the new custodian, as described below. As a final step, the old custodian removes any state associated with that user.

If a user has multiple custodians, then the sender, (or, for unbound bundles, the Internet gateway that discovers the set of custodians associated with the user) adds multiple custodian GUIDs in the bundle header of a bundle destined to that user, in addition to the user's GUID. The forwarding engine in the DTN router uses the custodian headers to deliver one copy of the bundle to each custodian named in the header. It is left to an application-layer protocol to delete (or to allow the time out of) bundles that have been delivered to the user, but are still pending delivery at one or more custodians.

Dealing with race conditions

Mobility intrinsically introduces race conditions. Bundles may be sent to one of the mobile's old custodians or local DTN router before it has heard of the mobile or after it has moved to a new custodian. Because these race conditions are not easily avoidable, we have designed our protocols to work correctly, though with reduced efficiency, in case of races. The basic design principles are to update location information *before* old information is deleted i.e. *make-then-break* and to be generous in accepting bundles. This way, bundles may be sent to an out-of-date location, but are very likely to be eventually delivered to the right destination.

Consider a mobile registered with custodian Old that moves to custodian New. On receiving a registration, custodian New first updates the HLR to point to itself, then tells Old that it is no longer the custodian. Bundles sent to the user after the HLR change will arrive at New correctly. However, bundles in flight may incorrectly arrive at Old. Worse, they may have been forwarded by Old on the path towards the user.

Bundles at Old that have not been picked up before the arrival of the custodian update message are forwarded to New when the update message arrives at Old. To sweep up bundles that have been forwarded to the user from Old, it forwards the custodian update message on the path to the user and every DTN router along the path resends these bundles to New by changing the custodian portion of the address to *unbound* and resending the bundle. Once this is done, the mobile user's state is removed from Old, and by every DTN router along the path from Old to the user. Subsequently, any bundles arriving to Old for that user will be bundles arriving to an unknown user, which we describe next.

Bundles that arrive to a custodian with a destination GUID that is unknown to that custodian (i.e do not have an entry in that custodian's local state) have arrived before the user has registered with the custodian, or after the custodian has cleaned up that user's state. In this case, the custodian first looks up the HLR to find the new custodian for that user. If the HLR has the new custodian for the user, then the custodian forwards the bundle to the new custodian. If the HLR has no information, then the custodian saves the bundle and awaits a registration until the bundle's time to live expires, with periodic HLR lookups to see if it can be handled by some other custodian.

3.4.2 Routing

The goal of a routing protocol is to map, at each DTN router, from a destination GUID (of a custodian or user) to the next hop link.

Getting bundles from a user or any DTN router to an Internet-based proxy is straightforward if we use default routing. Similar to the way routers are configured with a default route in the Internet, we also manually configure every DTN router with a default link which is the link on which to send a bundle whose destination custodian is *unbound*. For example, an end-point's default link would point to one of the ferries, which would have a default route to one of the Internet gateways.

Having a single default route is not fault-tolerant. To deal with failures, unbound bundles can be flooded in the disconnected region. This will lead to the same bundle being received by multiple gateways.Therefore, on receiving an unbound bundle, the gateway looks up the destination's GUID in the HLR to find its current proxy or gateway node, and then conducts a simple handshake protocol among other gateways in the disconnected region to avoid sending duplicate bundles to the same proxy or gateway. For small disconnected regions, we believe that flooding provides sufficient redundancy to achieve reasonably good performance in most cases.

The reverse path (i.e from proxy to kiosk) is much harder to determine. The proxy needs to know the best Internet gateway to use, and the Internet gateway needs to know how to reach the user's custodian. Finally, the custodian needs to know how to reach the user. Unfortunately, general routing in DTN is an unsolved problem. We therefore present three alternative routing protocols.

Flooding

This is the simplest routing scheme. Here, the proxy floods bundles into the entire DTN network. Reachability is guaranteed, but scalability is a problem, particularly given that the bottleneck link capacity, which is the access link from an Internet gateway, has a capacity of typically around 100kbps, or around 1 GB/day. Opportunistic links may also become a bottleneck because they are likely to have a maximum connection duration of only a few minutes, and flooding may result in excessive duplication of the same data on multiple links. So, flooding is unlikely to be useful in any but the smallest deployments.

Reverse path forwarding

This scheme is more efficient than flooding, but it gains efficiency at the cost of increased system fragility. In a nutshell, reverse path forwarding uses a locationing update to simultaneously set up forwarding tables along a sink tree, therefore combining locationing and routing. As we shall see, it also requires that custodians be either present at, or be associated with, a single Internet gateway, and, moreover, to be present on the path from the user to the Internet gateway.

In this scheme, the system uses only a single path to every entity in the system – all alternate paths are ignored. Therefore, the problem of reaching a user reduces to finding a unique path from a sender to every custodian for that user, and from every custodian to every user registered with that custodian.

We create paths to custodians by associating a unique Internet gateway with each custodian, and storing the Internet gateway's IP address in the HLR, in the same way as for Internet-accessible custodians. Unbound bundles associated with a particular custodian, therefore, automatically reach the Internet gateway as described earlier. The choice of Internet gateway for each custodian, and setup of a reverse path from the Internet gateway to a custodian and from a custodian to a user, are done using reverse path forwarding as described next.

Specifically, we define a special control message called the REGISTER message. When a mobile registers with a new local DTN router, or a new user is created at a kiosk, the user sends a REGISTER message containing the user's GUID and the GUID of its custodian, that is forwarded along the default path to an Internet gateway. Along the way, we require it to touch the custodian, therefore the custodian must lie on this path. As the REGISTER message propagates to an Internet gateway, all the DTN routers along the way record the previous hop of the message (i.e. the TCP/IP address of the previous DTN router) in their forwarding tables to be the next hop for any data addressed *to* the user. Thus, data to be transferred to the user follows the reverse of the default path taken by the REGISTER message, while data to be transferred to the Internet from the user follows the default path.

When an Internet gateway gets a REGISTER message, it updates the HLR to map the user's GUID to the GUID of the user's custodian as well as to its own IP address. All future communication with the user happens through that Internet gateway.

The same procedure is followed by custodians to select the Internet gateway they are associated with. Ferries also send REGISTER messages to discover the custodians reachable by them.

Reverse path forwarding is useful for routing bundles on shortcuts. That is, if the destination's GUID is known to a DTN router along the default path, then a reverse path exists to that destination, and bundles can be sent there directly. This allows, for example, a bus going from one village to another to carry bundles between the villages without having to go through the Internet. This allows rapid user-to-user communication without the mediation of a server.

The use of reverse path forwarding has both its pros and cons. On the one hand, in the absence of a definitive solution to the DTN routing problem, it offers a simple way to set up the routing tables in a network. On the other hand, these tables are fragile: if a link were to break, or a DTN router were to fail, the protocol does not recover gracefully from this failure. The lack of fault tolerance can be handled in several ways. For instance, a user or DTN router can periodically send a REGISTER message to refresh paths to it. These protocols would limit outages to approximately one update period, which may be sufficient in practice. Another disadvantage of this protocol is that it relies on manually configured default routes, and can result in sub-optimal performance. We are currently looking into how link state routing, described next, can be coupled with reverse path forwarding for better performance.

Link state routing

Link state routing has been proposed in recent work on practical DTN routing [33, 64]. In this work, standard link state packet flooding is used to construct network topology graphs, where link weights are set to the expected link delay. A shortest-path algorithm is then used to create forwarding tables. We believe that the solution described here can be used in our system (though as of now, our system only implements simple reverse path forwarding).

The problem of choosing link weights, in general, is a difficult one. The weights should represent not only bus schedules, but the proxy-gateway links should also take into account the queue-lengths at a proxy for data destined to different kiosks. This is because the proxy can reduce end-to-end delay by intelligently scheduling bundles on the proxy-gateway link ahead or behind of each other, depending upon the bus schedules from the gateways to the different kiosks. Essentially, the problem is complex because the time scale of data forwarding on the proxy-gateway link is the same as the time scale of routing. We do not yet have a complete solution to this problem, although a first attempt has recently been made in [62].

Note that even the bus schedules may not be very accurate because buses can get delayed or rerouted on a more or less random basis. Thus, the topology graph at each DTN router should be represented such that incremental changes can be made to it, based on local observations and periodic control plane updates. We are currently looking into adapting the work from [33, 64] into our network model, and extending it with routing and scheduling based on approximate information and incremental updates.

3.4.3 Application support

We would like to provide a simple API for application development that supports session persistence, intelligent use of multiple networks, and use of unmodified legacy servers. The Opportunistic Connection Management Protocol (OCMP) provides all these features. An OCMP client running on a mobile device can communicate opportunistically over multiple network interfaces to an Internet proxy. Legacy application protocols are hidden from the client through application-specific plugins that talk to legacy servers on behalf of the client.

OCMP differentiates between shared-control, application-control, and application-data packets. Shared-control packets are used primarily for connection state updates between the mobile device or kiosk-controller and the OCMP proxy (common to all applications); application-control packets are used for conveying application-specific parameters between the plugins on the client and proxy; and application-data packets are used for bulk data transfer. On the client side, an OCMP plugin is instantiated for every transaction (send or receive) by a legacy application. The plugin can create multiple data streams to the proxy, where each disconnection-tolerant stream is recognized by a unique session identifier. The proxy then reassembles the data of each stream, and passes it to a corresponding application plugin on the proxy side. This plugin can also receive application-control packets to reconstruct the application state required to instantiate a legacy transaction on behalf of the client. A similar procedure is followed for legacy communication from a server to the mobile device.

We extend OCMP to operate on an end-to-end basis even across multiple levels of disconnections. We do this by modeling DTN as a network interface for OCMP that implements its own transport layer protocol.

Whenever OCMP detects a WiFi network, it examines the SSID to determine whether the network belongs to a DTN router that is a part of our infrastructure,

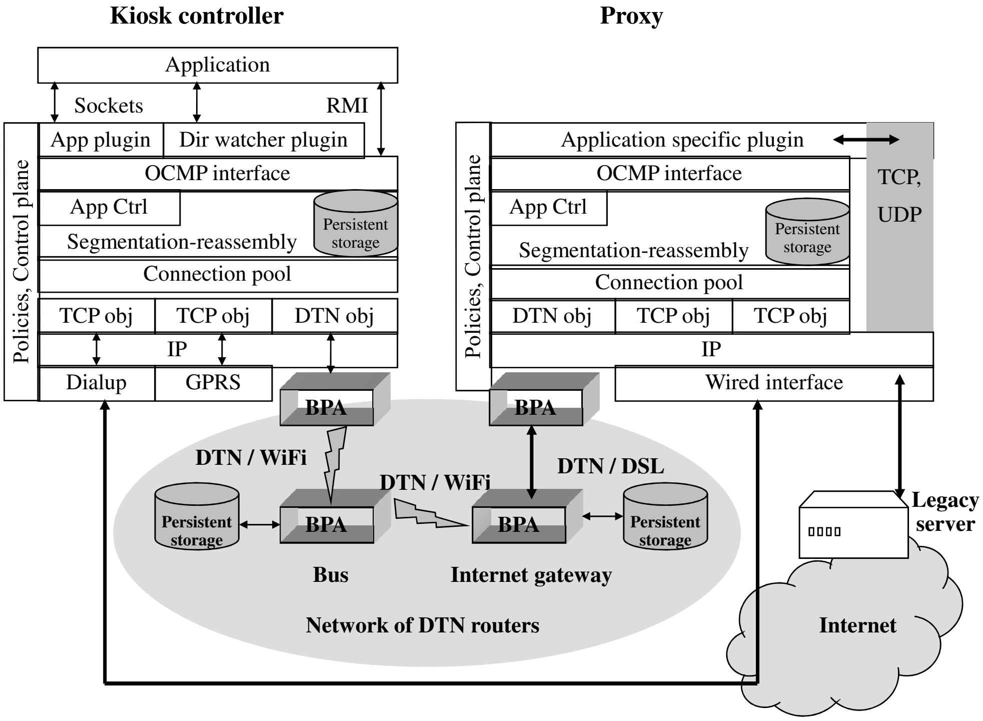


Figure 3.4: OCMP/DTN integration

or it is a third party WiFi network connecting into the Internet. A different type of connection is instantiated accordingly. Each connection instance in OCMP is encapsulated in a *Connection* object, and a collection of such objects is maintained in a *ConnectionPool* object. Therefore, we have implemented a new *Connection* object for DTN that talks to a DTN bundle protocol agent (BPA) running either locally or on a DTN router. This is shown in Fig.3.4, and discussed as follows.

OCMP on kiosk controller: A BPA runs locally on the kiosk controller. Thus, whenever OCMP detects a new network from a mobile DTN router, it opens a connection to the local BPA, and the BPA connects to the corresponding BPA on the mobile DTN router. OCMP then sends data to the local BPA which encapsulates it in DTN bundles and forwards the bundles to the correspondent BPA. Custody transfer acknowledgments are relayed back from the BPA to the DTN *Connection* object in OCMP. Thus, OCMP is made to believe that it is actually talking to a proxy in the Internet, but the BPA successfully masks the absence of an end-to-end route to the Internet. Similar steps are followed for data to be downloaded to the kiosk controller from the mobile DTN router. Note that the BPA passes data to OCMP only for kiosk users. For mobile users who use the kiosk controller only as a DTN router, this data is retained at the BPA itself.

OCMP on mobile host: A BPA does not run locally on the mobile host, but OCMP connects wirelessly to the BPA on the kiosk controller or a DTN router, and transfers data to it over the BPA's RPC interface. The subsequent working is identical to the previous case.

OCMP on proxy: For data to be sent to a kiosk controller or a mobile host, the proxy first checks whether the endpoint is registered with a custodian in the Internet region or not. This step is crucial for the proxy to decide whether to retain the data in OCMP or to push it into the DTN overlay by routing it to an appropriate Internet gateway. In the former case OCMP layers itself on TCP/IP and for the latter case, the proxy instantiates a connection to the BPA running locally and dispatches all the data to this BPA for eventual delivery to the user's custodian.

3.4.4 Opportunistic link use

We now present the sequence of actions that happen when an opportunistic connection is detected by a kiosk-controller or mobile device due to the arrival of a ferry. It serves to illustrate how the naming, addressing, and forwarding schemes come together at different layers of the protocol stack.

1. **Link association**: The WiFi NIC on the ferry uses active beaconing to broadcast its presence. The WiFi card on the kiosk controller detects the peer WiFi node when it comes in range, and attempts to associate with it.

2. **OCMP and BPA initiation**: OCMP running on the kiosk-controller detects a successful association, recognizes the SSID of the WiFi network, and notifies the BPA to connect to its correspondent BPA on the ferry.

3. **Data download**: The kiosk BPA requests the ferry BPA for any data addressed to the users registered at the kiosk. If data is available, it is downloaded and passed on to OCMP. OCMP reassembles the data and redirects it to appropriate application specific plugins.

4. **Data upload**: If there is upload data pending, OCMP dispatches its data to the BPA. The subsequent steps involved for data upload are similar to the download process, but in addition, we also allow the kiosk controller to decide which bundles will take which route, based on the routing protocol and current load conditions. This is done to accommodate load sensitive flow control in our architecture.

5. **Routing**: To enable routing based on load statistics, we allow the kiosk BPA to query the mobile DTN router about queue sizes and routing metric information, and source route the DTN bundles for upload accordingly. The ferry can continue to query other routers on the way, to decide optimal and load balanced routes for the bundles it is carrying.

6. **Session persistence**: OCMP application plugins can specify a session identifier for each data transaction like a file upload or download. This is used to identify the application endpoint at both the proxy as well as the kiosk controller to which data is finally redirected after reassembly. End-to-end session level ACKs are used to ensure that the data finally reaches its destination by incorporating a retransmission timeout in case some data gets lost in the DTN overlay due to system failures.

3.4.5 Scaling

With small sized deployments of the order of 10s of kiosks, reverse path forwarding routing on a flat topology may be adequate. The Indian government, however, has announced that it intends to deploy 100,000 kiosks across the country. Issues of scale, therefore, cannot be long ignored. Here, we sketch out an approach to allow scaling, deferring details to future work.

At first glance, the major non-scalable element in our architecture appears to be the use of non-aggregable GUIDs for routing. In fact, we do not believe this poses a significant problem. The HLR may need to track millions, or even billions of GUIDs. However, today's systems can handle gigabyte-sized tables in main memory, and clustered solutions that leverage Distributed Hash Tables can offer several orders more capacity and nearly perfect fault tolerance [68]. Similarly, routing tables in DTN routers can easily scale to tens of millions of entries because lookup delays of the order of milliseconds, or even seconds, are acceptable.

The real issue, we believe, is that a flat topology does not scale well with respect to the number of update messages. For instance, in reverse path forwarding, when a user moves from one kiosk to another, the new REGISTER message has to be sent all the way to the HLR. However, users are typically expected to move within villages close to each other. The standard technique to allow scaling is to introduce hierarchy. We now outline how we can add hierarchy in routing and location management to allow our system to expand to millions of users.

Essentially, we break up the topology into autonomous *regions*. A region is a collection of mutually reachable DTN routers, determined by administrative policies, communication protocols, naming conventions, or connection types. Regions would usually be contained within a physical boundary, though a single region may span multiple geographical areas. We do not place any assumptions on the organization of regions. From the perspective of the outside world, a region is reachable by a set of border gateways (much like BGP routers in the Internet).

To reduce the overhead in locationing, the HLR maintains a mapping from a user's GUID, not to a custodian, but to the *border gateways* of the region in which the user is present. Additionally, a region's border gateways maintain a replicated copy of a VLR (Visitor Location Register), that carries a mapping from the user's GUID to its current custodian. Therefore, if a user changes custodians, but stays within a region, it does not need to update the HLR, increasing scalability.

Region partitioning also improves the scalability of routing. Suppose link state routing is being used. In this case, routing updates within a specific region need not include information of the entire network, but only the intra-region routing state of that region. This significantly reduces the number of routing updates. Similarly, if flooding is used, instead of flooding the entire network, only the destination region needs to be flooded.

3.4.6 User-to-user vs. User-to-server communication

We now discuss one aspect of our architecture that is non-obvious, but relevant to our overall goals, that is, its support for user-to-user communication. In other words, our system allows a user to directly send data to another user without having to go through an intermediate server or proxy. The significance of this decision has to do with efficient communication paths between users to enable the distributed recommender system architecture discussed later in Chapter 8.

Consider a user in a village who wants to send a message to another user in one of the villages served by the same ferry. This is likely to be common because most communication is local. In a system that supports only user-to-proxy or user-to-server communication, the sender would send a message addressed to the receiver, but tunneled to a well-known proxy or server. This forces data to traverse the path to the proxy or server and back. This is just another version of the *triangle routing* problem in Mobile IP [67]. To avoid this, it is necessary to permit direct user-to-user communication. This requires public key information and locationing information to be disseminated in the network. Our use of hashes of user names as forwarding identifiers, and using the same identifiers for public keys, allows us to integrate naming, addressing, routing, and security.

3.5 Deployment

We deployed an early version of the system in May 2006 in Anandpuram, a village in South India about 20 km from the city of Vishakapatnam. The Vishakapatnam District Rural Development Agency (DRDA) has set up over 40 kiosks around Vishakapatnam, and the Anandpuram kiosk is part of this initiative. At the kiosk, we set up a recycled PIII PC as a public access terminal, and connected the kiosk PC and the recycled PC with a Net4801 Soekris Single Board Computer (SBC) as the kiosk controller. We used a similar SBC as the Internet gateway with DSL broadband (128 kbps) at the DRDA head office in Vishakapatnam. At both places, we used external 9 dBi omni-directional antennas for wireless connectivity, and 40 GB hard-disks for local storage. The kiosk SBC runs from a 42 AH battery that is charged by two 1.2A @ 12V solar panels, and the gateway SBC runs on UPS. This ensures 24 hour uptime for both the nodes. For the ferry, we chose a government vehicle that regularly goes between the head-office and the kiosk for a microfinance initiative also led by the DRDA. We installed a similar Net4801 SBC in the vehicle, powered it from the vehicle battery, padded it with foam to protect it from vibrations, and attached an external 7 dBi omni-directional antenna with a magnetic mount base that sits outside on the trunk. Finally, we integrated the e-governance portal with the OCMP API to allow for delay tolerant connectivity.

Further development of the system has been undertaken by other members in our research group, and has resulted in more deployments.

3.6 Discussion and related work

We draw upon a wealth of ideas in the literature that deal with disconnection tolerance, mobility management, semantic-free naming, data privacy, and routing. We have outlined the principles derived in past work in Section 3.1. Here, we consider systems that are most closely related to ours.

The use of data ferries in the context of MANETs and sensor networks is well known; for example, see [70]. Current work on DieselNet [59] is also relevant. However, these systems essentially present point solutions that only address a few of the goals outlined in Section 3.1. Moreover, this work is not directly applicable to low-cost and reliable kiosk networking in rural areas.

The work closest to ours in spirit is that of Daknet [66]. They use MAPs (Mobile Access Points) mounted on buses or vans, which regularly traverse villages and come in wireless contact with rural kiosks to opportunistically upload and download data. Buses were fitted with omnidirectional antennas, and kiosks with omnidirectional or directional antennas depending upon the orientation of the kiosk with the road. Data sessions of an average duration of 2:34 minutes were measured, during which up to 20 MB of data could be transferred. We have experienced similar performance because our operating environment is practically the same as that of Daknet. However, Daknet does not implement a generic architecture that can be used to build new applications, support integrated location management system, operate across multiple levels of disconnections, or allow (to our knowledge) data privacy. Our system provides all these features, along with many rich applications ready for immediate use.

Our work relates broadly to rendezvous-based mobility support in HMIP [69] or I3 [34], location-independent identifiers in HIP [40], and semantic-free names in DoA (Delegation Oriented Architecture) [58], but these protocols are designed to work in connected Internet like environments. We have suitably adapted the insights from these schemes into our architecture.

DTN [71] also forms a fundamental part of our architecture. However, DTN is a general platform, and we have specialized and extended it to handle our specific interest.

Chapter 4

Network security for KioskNet

The goal of computer security: Computers are as secure as real world systems, and people believe it – Butler W. Lampson, Computer Security in the Real World, 2004

In the previous two chapters, we described how delay tolerant network design can reduce the cost of connectivity for mobile phones and rural kiosks in both user-to-user and user-to-server communication scenarios. Support for these connectivity patterns will be needed to enable the distributed recommender system design discussed in Chapter 8. The provision of security in such network designs is however a daunting task because traditional mechanisms such as provisioning a Public Key Infrastructure (PKI) are not well suited to environments where nodes may be disconnected for long periods of time and end-to-end communication is usually not possible.

Two problems that arise are (1) establishing a secure channel between a sender and receiver separated by multiple hops of disconnection, and (2) mutual authentication between a pair of nodes at either end of an opportunistic link. With a PKI, a sender can establish a secure channel by encrypting data with a session key, and encrypting the session key in turn with the recipient's public key. However, a disconnected sender cannot efficiently use a PKI because finding out a recipient's latest public key requires an end-to-end round trip to a central or replicated lookup database, substantially delaying actual data transmission. Similarly, mutual authentication can be assured through a PKI by means of certificates issued to both endpoints of an opportunistic link by a mutually trusted third party. However, this requires authenticating parties to carry certificates from the same trusted authorities, which may not be likely in the kind of rural connectivity scenarios we consider.

Our contribution is the development of a practical cryptosystem for disconnected environments using Hierarchical Identity Based Cryptography (HIBC) [73] for creating secure channels and providing mutual authentication, even across different administrative domains. We have also developed procedures for initial key establishment to allow new users to sign-up into the system in a secure manner, and a simple technique to prevent a user's identity from being compromised due to the loss or theft of a mobile device. Our solution is novel in that it explores the practical aspects related to deployment of delay tolerant networks. In this chapter, we describe our solution for security in the context of the mechanical backhaul system for rural kiosk connectivity explained in Chapter 3, but our solution is easily generalizable to OCMP (Chapter 2), which is only a special case for mechanical backhaul with zero or one stationary disconnection hop.

We start by presenting a review of the mechanical backhaul topology design described in the previous chapter, then define the threat model for our security solution in Section 4.2, and give details of the security schemes in Sections 4.3 and Section 4.4. Finally, related work and discussions are presented in Section 4.5.

4.1 Mechanical backhaul overview

We present some definitions first introduced in Chapter 3.

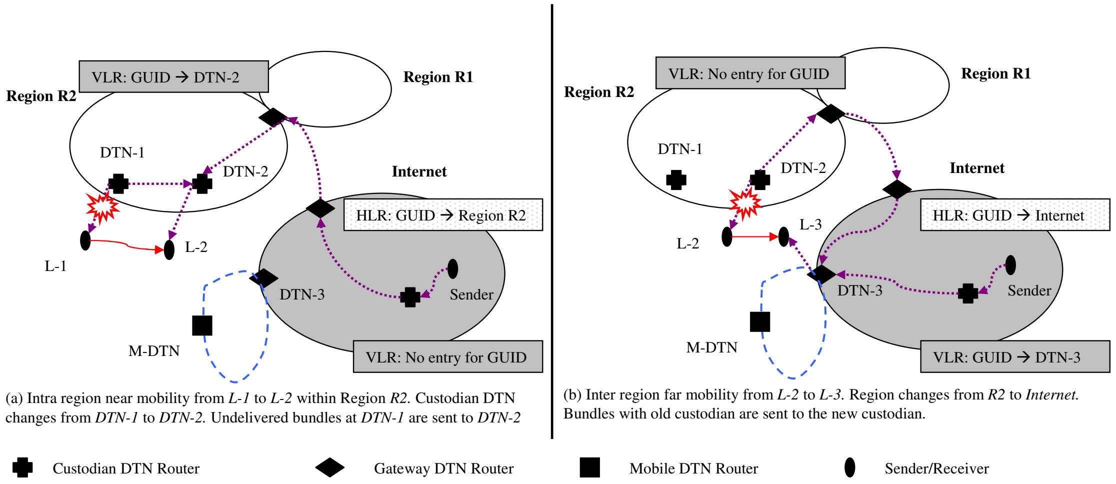
1. *Region*: A region is a collection of mutually reachable DTN routers, determined by administrative policies, communication protocols, naming conventions, or connection types. The Internet is a single DTN region.

2. *Gateway*: This is a DTN router with interfaces on more than one region. An Internet gateway is a DTN router with at least one interface to the Internet region.

3. *Custodian*: This is a DTN router that acts as always-available proxy for intermittently connected hosts. Custodians opportunistically receive bundles from disconnected hosts, forward them to other custodians, and deliver them to a receiver whenever the receiver connects to the network.

4. *Mobile DTN router*: This is the DTN router that communicates directly with an endpoint. A mobile DTN router may or may not be a custodian as well.

Consider the following scenario. A PDA user in a village without any Internet connectivity offloads all her data into a village kiosk. When a bus drives past this Internet kiosk, it picks up the data locally stored at the kiosk, and also picks up data from other PDA users and kiosks on the same bus route. The bus then enters into a city and offloads all its data into an identical kiosk located at the bus station. This kiosk is connected to the Internet over a slow DSL connection, and uploads all this data into a proxy. The proxy reassembles the data and dispatches it to legacy servers like email or content servers. In this scenario, the village kiosk is the custodian DTN router for PDA users, the bus is a mobile DTN router, and the bus station kiosk is a gateway between the village region and the Internet region.



(a) Intra region near mobility from *L-1* to *L-2* within Region *R2*. Custodian DTN changes from *DTN-1* to *DTN-2*. Undelivered bundles at *DTN-1* are sent to *DTN-2*

(b) Inter region far mobility from *L-2* to *L-3*. Region changes from *R2* to *Internet*. Bundles with old custodian are sent to the new custodian.

Figure 4.1: Lookup hierarchy

The bus station kiosk and the village kiosk may or may not be present in the same region. In either case, the mobile DTN router ferries data between the two kiosks.

Fig.4.1 shows a mobility scenario for mobile users who travel between different kiosks in a region or across regions. We assume that all mobile users are identified using an opaque globally unique identifier (GUID). The Internet region maintains a registry called the Home Location Register (HLR) that maps the mobile's GUID (I) to its current region (R). Each region maintains a Visitor Location Register (VLR) that stores a mapping and path from the GUIDs (I) of all mobiles currently in the region to the mobiles' custodian DTN routers (C). Finally, each custodian maintains a Local Location Register (LLR) that maps from the GUID to the best last-hop fixed or mobile DTN router (M) for each mobile.

We assume that every DTN router has a default entry in its routing table that allows bundles to be forwarded (eventually) to an Internet gateway. Additional entries in the routing tables in a region are established using reverse-path-forwarding when a REGISTER message sent by a user's device propagates towards its closest Internet gateway along default pre-established routes. When the mobile device moves, its location information is updated in the appropriate location registers using a new REGISTER message. Outdated entries are either deleted or automatically time out.

4.2 Threat model

We assume the following threat model:

1. Eavesdroppers can potentially overhear (wireless) communication.

2. Rogue DTN routers that are not a part of the infrastructure, may pretend to be valid DTN nodes and attract traffic for which they are not responsible.
3. Malicious users may inject undesirable traffic into the DTN infrastructure in a DoS attack.

Given this threat model, consider a user who would like to conduct a secure transaction using the mechanical backhaul system. Because infrastructure nodes cannot be trusted, every opportunistic link must include a phase of mutual authentication between the two endpoints. Second, users will want to establish end-to-end secure channels to prevent eavesdropping. Third, the infrastructure must protect itself from malicious users who may try to launch DoS attacks. This requires techniques to establish end-to-end secure channels, perform mutual authentication, and maintain audit trails to detect malicious users. We now describe these mechanisms after giving an overview of HIBC.

4.3 Hierarchical Identity Based Cryptography

Boneh and Franklin [74] proposed the first practical Identity Based Cryptography (IBC) scheme and many variations have subsequently been described in the literature. Unlike traditional PKI, where a user obtains a certificate for her public key pair from a certifying authority, public keys in IBC can be any string, but private keys are obtained from a trusted authority called the Private Key Generator (PKG). Thus, possession of a valid IBC private key implicitly implies certification as well. Hierarchical IBC extends IBC by establishing a cooperative hierarchy of PKGs. The top-level PKG is called the root PKG, and the other PKGs are called domain PKGs, each of which inherits the first part of its public ID from its parent. We represent the public key of a user at level t in the key hierarchy as $(ID_1...ID_{t-1}.Username)$, where $(username)$ and (ID_1), (ID_1, ID_2), ... $(ID_1...ID_{t-1})$ are arbitrary strings for the identity of the user and domain PKGs along the hierarchy. This indicates that the parent PKG of the user is the domain PKG at level $t-1$ with a public identifier of $(ID_1...ID_{t-1})$.

Identity Based Cryptography (IBC) is ideally suited for creating a secure channel in a disconnected environment because the public key of an entity can simply be its public ID, and hence a lookup step is not required. For example, the public ID for a user can be the email address of the user herself. Another advantage is that the possession of a valid private key implies that the certification authority has certified the identity. Therefore, a valid signature serves as an assurance of authentication. Finally, a user can freely generate certificates signed with her private key that are universally verifiable.

A detailed description of Hierarchical Identity Based Encryption (HIDE) and Hierarchical Identity Based Signature (HIDS) is given in [73]. The root PKG and each domain level PKG share a set of publicly known system parameters, and a secret parameter local to each PKG. A combination of the secret parameter and system parameters are used to generate a unique private key for every public key that is requested.

It is well known that HIBC suffers from lack of forward secrecy, meaning that the compromise of a key can compromise earlier transactions as well, which were conducted using that key. A forward-secure HIBC scheme is proposed in [77]. We propose a simpler algorithm for forward-secure HIBC in Section 4.4.2.

HIBC also suffers from the problem that if a PKG is compromised then it can yield all the private keys generated for lower level PKGs and users, which can be maliciously used to decrypt messages. A combination of IBC and PKI [76] has been shown to avoid this problem, but we cannot use this scheme because it does not allow the public key to be chosen freely. Rather, the public key of a user is derived from the IBC portion of her private key. In our approach, we assume that the PKG nodes are trusted and cannot be compromised. We are investigating alternative solutions to this problem in ongoing work.

4.4 Security mechanisms

4.4.1 Establishing secure channels

We incorporate HIBC by assuming that a tree-like hierarchy can be imposed on the DTN regions, based on administrative structures and policies. Thus, each provider maintains its own top-level PKG, preferably in its partition of the DTN Internet region. Every sub-region has its own domain-level PKG; alternatively, location registers in the sub-region should be able to default-route key pair requests to a parent PKG. A user can request a public ID and private key either from his nearest regional PKG or directly from the top-level PKG. The procedure of acquiring public-private key pairs is explained in Section 4.4.2, and needs to be executed only once for new users who need a DTN identity. Each DTN router also maintains a unique identity for itself granted by the PKG of the region to which the DTN router belongs.

HIBC now allows the creation of an end-to-end secure channel: the sender encrypts all data with the public key of the recipient, and only the recipient can decrypt the data. The sender also signs the data, and the signature is verified by the receiver. This provides confidentiality, integrity, and authenticated access. Besides allowing end-to-end secure channels, HIBC also protects the infrastructure from a class of attacks on the location management subsystem. Recall that a mobile device sends control messages whenever it changes its location. We use HIDS between the mobile device and the location registers being updated, with the system parameters of the mobile device's HIBC system piggybacked on the message. This ensures safety from fabrication of control messages, redirection attacks, and the creation of dead-ends by unauthorized updating of location registers. Finally,

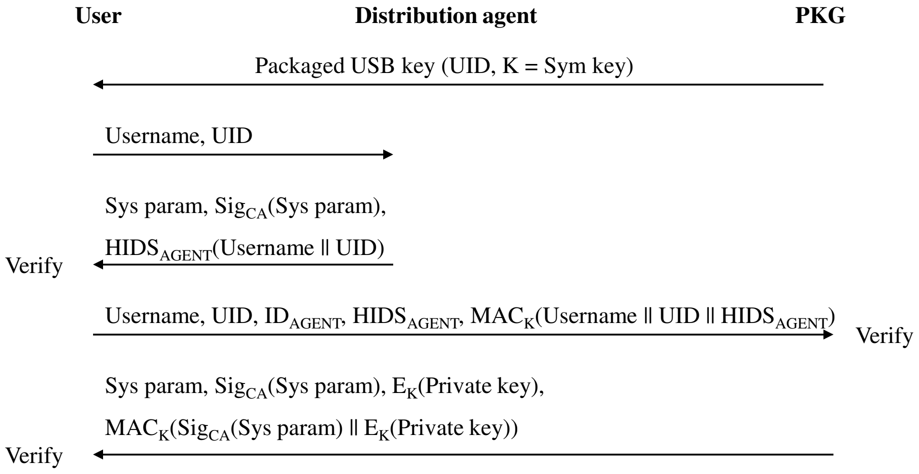

User
Distribution agent
PKG
Packaged USB key (UID, K = Sym key)
Username, UID
Sys param, Sig_{CA}(Sys param),
HIDS_{AGENT}(Username || UID)
Verify
Username, UID, ID_{AGENT}, HIDS_{AGENT}, MAC_K(Username || UID || HIDS_{AGENT})
Verify
Sys param, Sig_{CA}(Sys param), E_K(Private key),
MAC_K(Sig_{CA}(Sys param) || E_K(Private key))
Verify

Figure 4.2: Establishment of system parameters

DTN custodians send messages to the mobile device whenever they undertake the custody for a data bundle. We require custodian DTN nodes to sign these messages for the bundles that they take custody of, in order to ensure safety from spoofing. Mobile devices can even store these acknowledgements for auditing purposes. Since the HIDS scheme itself ensures non-repudiation, the audit logs can be used as proof of custody transfer.

4.4.2 New user sign-up

A new user who can directly connect to the PKG can obtain its private/public key pair by communicating with the PKG over a standard secure channel mechanism like SSL. However, if the new user is in a disconnected region, it cannot communicate with the PKG. How then should it obtain its keys? We show this process in Fig.4.2. We propose that USB storage devices (such as the popular 'USB keys') be used by the PKG to distribute keys through authorized distribution agents to disconnected end users. For instance, these pre-loaded USB keys could be given to a kiosk operator who authenticates a user first-hand and then hands over a USB key (similar to the way **SIM cards** are handed out for cell phones today). These storage devices carry a pair of (*UID*, *Symmetric key*) that has been generated by the PKG. During the setup phase, mobile hosts send their desired username and UID to the distribution agent. The agent signs the tuple, and sends back to the user the signature along with the system parameters of the HIBC scheme being used. The system parameters are themselves signed by a well-known certifying authority like Verisign to ensure the authenticity of the provider. The user verifies the signatures to confirm that the provider is real, and the agent has matched the desired username with the UID. The user then authenticates the signature, username, and UID by a MAC on the symmetric key, and addresses it to the PKG.

The PKG looks up this UID to verify the MAC, and uses the requested username to determine the user's public key. It then computes the private key for the mobile host and sends it back to the user encrypted on the same symmetric key. Because the symmetric key is a one-time secret shared only by the new user and the PKG, this assures the security of this communication. To prevent the kiosk operator or distribution agents from tampering with the USB storage device, the device itself can be wrapped in a tamper-resistant package (such as a sealed cellophane wrapper), which can be verified by visual inspection.

Note that because we require an authorized agent to distribute the USB keys, we assume that if a user desires to use a well-known ID like his email address as his public ID, then the authorized agent can verify that the email address being requested by the user is indeed the user's own email address. Second, also note that a new user is actually unreachable because no location table entries exist for that node's UID! How can the PKG send a reply to this user? If we allowed temporarily unverified entries in the location registers, we would open a security hole that could be used for a DoS attack. So, for this special case, we use source routing. Specifically, when the new user's message is sent to the PKG, it accumulates a certified route as it propagates towards the PKG. The PKG simply reverses this route and source routes the reply back. This allows the new user to be added to the network without trust violations. Once the user is added, it can register its location with a signed REGISTER message.

Preventing identity theft due to loss of mobile devices

Users will generally access the DTN infrastructure through mobile devices like cell phones and PDAs. However, such devices can be easily lost, which also implies a loss of the identity. Our solution is to never keep the actual private key on the PDA, but to extend the key hierarchy by another level that is time-based. In other words, public and private keys for $(ID_1...ID_{t-1}.Username)$ are extended to $(ID_1...ID_{t-1}.Username.Date)$. Here, the user acts as a PKG for herself, and can generate and expire keys at any desired time granularity based on her mobility and connectivity profile.

These time-based keys are generated for each new day in a secure location, such as on a desktop, and downloaded to the PDA periodically, say every few days. Thus, even if the PDA is lost, an intruder will gain access for only a limited amount of time (i.e until the keys on the PDA remain valid). Even this access can be prevented by prompt notification to the infrastructure through out-of-band mechanisms to quarantine all resources belonging to the user till the time-based keys have expired. The advantage with time-based keys is that the duration of the compromise is limited. Frequent updates can bring down the duration of exposure to arbitrarily small values, taking user mobility patterns into account. We believe this is an adequate practical solution for forward secrecy in IBC systems.

4.4.3 Mutual authentication

A typical communication session in the system involves the following steps:

1. *Initial setup*: When a mobile device desires to become a member of the DTN system hosted by a certain provider, it acquires the software and the security parameters from the resellers or distribution agents of that provider, as explained in the previous section.

2. *Link association*: When a mobile device and a mobile or fixed DTN router establish an opportunistic wireless connection, the mobile tries to connect to the router on a well-known port. Similarly, whenever a mobile device acquires a connection into a public network, it tries to connect to its DTN custodian or gateway.

3. *Mutual authentication*: Once a connection has been established between the mobile device and the DTN router, both participate in a mutual authentication procedure to ensure that malicious users and rogue DTN routers are not involved in the communication session.

4. *Location management and routing*: DTN routers carry authenticated routing and location management information from DTN custodians or gateways, which they give to mobile devices. The mobiles then use application specific policies to select their custodians or gateways, and send appropriate control messages to the DTN routers.

5. *Data transfer and billing*: Subscription plans can be created to enable data based billing. Thus, both mobile hosts and DTN routers collect mutually authenticated statistics on the amount of data transferred between the two entities. These statistics are non-repudiable, and can be verified later if the need arises.

6. *Roaming access*: Much like roaming access provided across different cellular networks, it can be enabled across DTNs hosted by different providers. Mobile devices and DTN routers exchange system parameters to be able to communicate with each other, along with collection of verifiable billing statistics.

We next describe the details for steps 3, 5, and 6.

Challenge-Response

If the DTN router belongs to the same provider as that of the user, a 1.5 round-trip challenge-response protocol is used to verify the authenticity of the user and the DTN router [80], shown in Fig.4.3. Here, R_1 and R_2 are random nonce generated as challenges by either party.

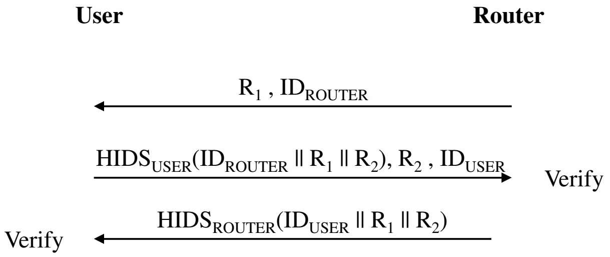

User
Router
R1 , IDROUTER
HIDSUSER(IDROUTER || R1 || R2), R2 , IDUSER
Verify
HIDSROUTER(IDUSER || R1 || R2)
Verify

Figure 4.3: Mutual authentication challenge-response protocol

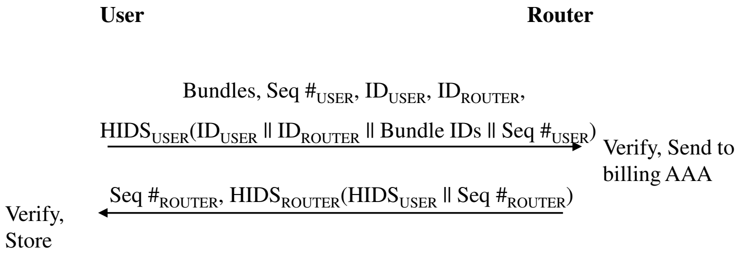

User
Router
Bundles, Seq #USER, IDUSER, IDROUTER,
HIDSUSER(IDUSER || IDROUTER || Bundle IDs || Seq #USER)
Verify, Send to billing AAA
Seq #ROUTER, HIDSROUTER(HIDSUSER || Seq #ROUTER)
Verify, Store

Figure 4.4: Data based billing

This protocol verifies that both the user and the DTN router are who they claim to be, and they possess valid private keys because they are able to sign each other's random messages. The DTN infrastructure is assumed to belong to the Trusted Computing Base (TCB), and hence it relies on the ingress DTN router to authenticate the user. Per-hop authentication can also be done if there is a high risk of the ingress routers to get compromised, and to push fake traffic into the network. In addition, all DTN nodes are assigned a "DTN" prefix in their public ID string, which can be used as an additional safeguard.

Billing

Data based billing is supported by negotiation of signed tokens between a user and the DTN router, before either party commits to sending or receiving data bundles. These tokens are sent to the DTN provider's billing server to bill the user. Incorrect billing is avoided due to the non-repudiation properties of HIBC. The detailed procedure is as follows, and illustrated in Fig.4.4.

1. Authentication tokens are created and signed by users through HIDS for each bundle or group of bundles. These tokens contain the user identifier and the identifier number of the bundles. The signed tokens and the bundles are sent by the end-host to the local DTN router.

2. The router verifies that the identifier numbers of all the bundles are included in the token, and sends a signed acknowledgement back to the user.

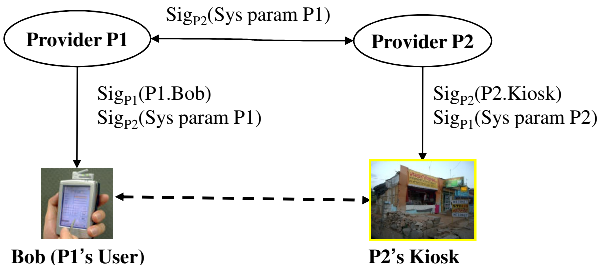


Figure 4.5: Chain of trust

3. The user stores the signed acknowledgement for auditing purposes to detect incorrect billing.

4. The router dispatches the data to the desired destinations, and sends the signed tokens to the DTN provider's billing server in order to impose charges on the user.

5. The signed tokens prevent any tampering attacks before presenting the tokens to the billing server. To avoid replay attacks, all authentication tokens carry the sequence number specified by the user as well. The billing server has to be careful in collecting and ordering the tokens according to the sequence number before analyzing them.

This solution is also privacy friendly since the billing server is given no knowledge about the contents of a bundle.

Roaming access

The DTN router can also belong to some other provider referred to as the roaming provider. This can occur if a bus drives past a PDA in a remote region, or if the PDA is taken into a remote kiosk, where the bus and the kiosk belong to a provider other than the home provider of the user. There are two cases in such a scenario:

1. **If guest access is allowed:** We explain this through an example illustrated in Fig.4.5 that uses the notion of chains of trust [75]. It illustrates a scenario where Bob, who is a user of provider P_1, roams to access service from a kiosk that is owned by provider P_2. To authenticate Bob, the kiosk is expected to possess P_1's system parameters signed by P_2, and receive Bob's public key signed by P_1. Since the kiosk trusts P_2, it infers that P_2 has allowed access to P_1's users because P_2 signed the system parameters of P_1. Now, since Bob is a valid P_1 user, hence the kiosk grants access to Bob through the chain of trust. Similarly, Bob verifies that he can trust P_2's kiosks by looking at P_2's system parameters signed by P_1, and the kiosk's public key signed by P_2. This shows that our scheme works well despite the entities being disconnected from each other. Note that HIBC is not necessarily required for this scheme,

and PKI is usable as well. We assume that any required signed parameters are available with the entities because the parameters are public and can be downloaded beforehand.

We slightly modify the first two steps in the 1.5 RTT challenge-response protocol shown in Fig.4.3 for mutual authentication in the single provider case. As explained in the example above, along with their respective public keys, both entities also furnish their respective system parameters signed by a well-known signing authority like Verisign, or signed by the correspondent entity's home provider. Once the system parameters have been negotiated and verified, the same protocol can be used to authenticate both parties.

Billing is done in a similar fashion by extending the scheme described in the previous section. Both users and DTN routers also include the identifier of their respective providers in the token signatures. Each entity signs the tokens in its own HIBC system, but verifies the tokens sent by the other entity in that entity's HIBC system. This is possible because signature verification does not require both entities to belong to the same HIBC system, but only requires knowledge of the public system parameters. This method can support *guest access* to allow data based post-payment of roaming services.

Note that the above mechanism can be used only for sending data through a roaming provider's network. However, if data needs to be received in a roaming provider's network as well, then appropriate routing tables need to be set up in the roaming HLR, VLRs, and LLR as well. This can be achieved in two ways. (a) The user signs the control messages through HIDS, and also attaches the system parameters of their HIBC system to the messages so that DTN routers along the route can verify the message authenticity. (b) The user is granted a temporary time-based identity by the local DTN router in the roaming network, similar to the roaming token method explained next. In either case, the home HLR of the user redirects all incoming data requests to the roaming HLR, much like the way a mobile IP home agent operates. Note that if control message encryption is desired as well, then only the second method can be used.

2. **If roaming tokens are granted in advance:** In this scheme, the soon-to-be-mobile user is given, in advance, a time-based private key and system parameters of a roaming provider. This can be done beforehand with secure communication between the user and the home provider, and the home provider and the roaming provider. The same 1.5 RTT challenge-response protocol is then used for authentication purposes between the roaming user and the local DTN router of the third-party provider. The temporary identity is used to set up routing tables in the roaming provider's network, so that the user can even receive data in the roaming network. The roaming token method is meant to support data based pre-payment of roaming services.

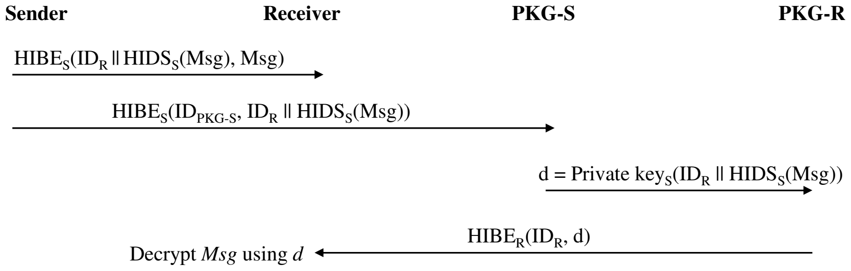

Sender
Receiver
PKG-S
PKG-R
HIBE$_S$(ID$_R$ || HIDS$_S$(Msg), Msg)
HIBE$_S$(ID$_{PKG-S}$, ID$_R$ || HIDS$_S$(Msg))
d = Private key$_S$(ID$_R$ || HIDS$_S$(Msg))
HIBE$_R$(ID$_R$, d)
Decrypt *Msg* using *d*

Figure 4.6: Cross-domain secure communication: $HIBE_S(e, M)$ stands for encrypting M on e in the HIBC system of S

4.4.4 Cross-domain secure communication

We have so far assumed end-to-end communication to exist between users of the same provider. Thus, the sender encrypts data on the public key of the receiver, and the receiver decrypts the data by its private key. However, if the receiver belongs to some other DTN provider, then the sender will have to fetch the system parameters of the receiver's HIBC system and encrypt the data using these parameters. This additional step to fetch the system parameters defeats the purpose of using IBC because it will involve a round-trip to the Internet. The parameter lookup will not be required if the ingress DTN router carries the system parameters of the most popular DTN providers. Here, we explain the procedure in the case when the receiver's system parameters are not available readily with the ingress DTN router. This is shown in Fig.4.6, and explained as follows.

1. The sender encrypts data in its own HIBC system by deriving a temporary session key to be used as a public key, to reduce the possibility of replay attacks. This key can be based on a monotonically increasing sequence number, or as shown in Fig.4.6, it can be uniquely based on the message signature itself. The encrypted data is sent directly to the receiver, but the receiver cannot decrypt the data until it possesses the corresponding private key. Note, that the data path for the encrypted message from the sender to the receiver can be derived through hot-potato like routing algorithms [81] that can work in a disconnected manner.

2. The private key for decrypting the message is generated by the PKG of the sender, and sent securely to the PKG of the receiver. Presumably, both of the PKGs will be present in the Internet, and therefore SSL-like mechanisms can be used to securely exchange information. Note that any other trusted entity could also be used here, instead of the PKGs.

3. The receiver PKG then encrypts the private key and sends it to the receiver in the normal manner in which it would send any kind of control message

updates to the receiver.

Thus, encrypted data is sent directly to the receiver, while the decryption key is derived by the PKG of the sender and sent to the receiver. This enables secure communication even between users of different DTN providers.

4.5 Discussion and related work

Security in disconnected environments has only recently been studied in the literature. Related work includes:

HIBC for DTN: The use of HIBC for DTN as well as the use of time-based keys for access control was proposed by Fall [82]. In contrast, we provide practical schemes for key dissemination, mutual authentication, secure location updates, audited custody transfer, and time-based forward secrecy.

PKI for DTN: All the security protocols described in this chapter will also work with PKI instead of IBC, as long as the public keys of recipients are known to the senders. This can be done through the broadcast all public keys on the network, but the method does not scale.

Offline authorization frameworks for ubiquitous computing: The 'Lobby' system proposed in [79] is typical of secure ubiquitous computing architectures like UPnP and Jini. These architectures provide an authorization framework that support offline mobile devices. These schemes require all policy-enforcement-points (or 'Lobby's) to periodically connect to a central database and renew their user and role based ACLs. Our work supplements such architectures by providing a general platform over which fine-grained trust models can be built to provide policy-based access control at the ingress points to the DTN TCB.

Anonymized communication in a DTN: [83] proposed an enhanced version of our scheme that also ensures anonymity in disconnected communication environments. A signcryption variant of IBC is used where two users belonging to the same PKG can non-interactively compute a shared secret, given their own private key and the identity of the correspondent user. Senders can then obfuscate their identity through a random nonce to provide sender anonymity.

Kiosk security: [84] did a detailed threat analysis of rural kiosks assuming that the kiosk operator cannot be trusted, and can snoop into users data stored on the kiosk controller. A detailed security architecture was then presented under this new threat model, and has been implemented as a part of KioskNet. Our design for network security discussed in this chapter supplements their work by proposing mutual authentication, billing, and roaming protocols. However, [84] chose to use PKI instead of IBC because of the licensing rules of the IBC implementation provided by Voltage Inc., the only vendor, which prevent Voltage's proprietary software from being used in foreign countries such as India.

Chapter 5

Message usefulness: An overview

The medium is the metaphor – Neil Postman, Amusing Ourselves to Death, 1985

Recall that in Chapter 1, we outlined two broad goals for the thesis: design of a communication infrastructure to provide *low-cost* and *universal* access to users, and a recommender system to push *useful* information using the communication infrastructure. Chapters 2,3, and 4 addressed the first goal through a discussion and prototype of the communication infrastructure, along with a proposal to ensure security using identity based cryptography. This chapter, and the next 3 chapters, address the second goal of designing a recommender system to determine what information to provide to users.

Our approach for providing messages to users is focussed on examining news-related items that exist within participatory media environments. We present in this chapter a perspective on participatory media, and an overview of how we propose to recommend messages to users in this setting.

Participatory media is distinguished from traditional forms of mass-media in that users participate to create and enhance media content, as opposed to content creation by centralized agencies. It has come to include blogs, discussion forums, wikis, tagging, and more such "Web 2.0" mechanisms which facilitate users to participate in content enhancement [85]. We specifically focus on blogs and discussion forums as forms of participatory media that allow users to articulate their opinion or provide facts, and are published on the Internet to remain visible to other users. Thus, we do not focus on wikis which require a consensus among users on content

Table 5.1: Characterization of media delivery mechanisms

	One-to-one	**One-to-many**
Unidirectional	telegraph	television, radio, books
Bidirectional	cellphones, letters	blogs

that is published, or on tagging which is used as a user-defined classification mechanism. Although our insights and findings can be extended to user participation in photo- and video-sharing systems, for ease of exposition we focus only on textual content in this thesis. Table 5.1 characterizes various information exchange mechanisms in terms of whether they enable *one-to-one* or *one-to-many* information transfer between people, and whether the channels support *unidirectional* or *bidirectional* communication. Blogs fall into a new category that was not available on a mass-scale ever before in human history. The closest equivalent of blogs seem to be the 18^{th} century *salon* (coffee-shop) discussions that served as a venue for people to engage in conversations with each other about politics and society [86]. Blogs respect the same conversational spirit of argumentation and opinion articulation; the participatory-Internet may indeed be the modern salon for human society.

Taking the salon metaphor forward, since participation in the salon discussions was open to anybody, the discussions helped people from different contexts to understand each other and form a more complete picture about various matters. Participatory media seems to play the same role by helping people to develop a better understanding of the complete picture. We therefore model participatory messages in terms of two types of information provided by them:

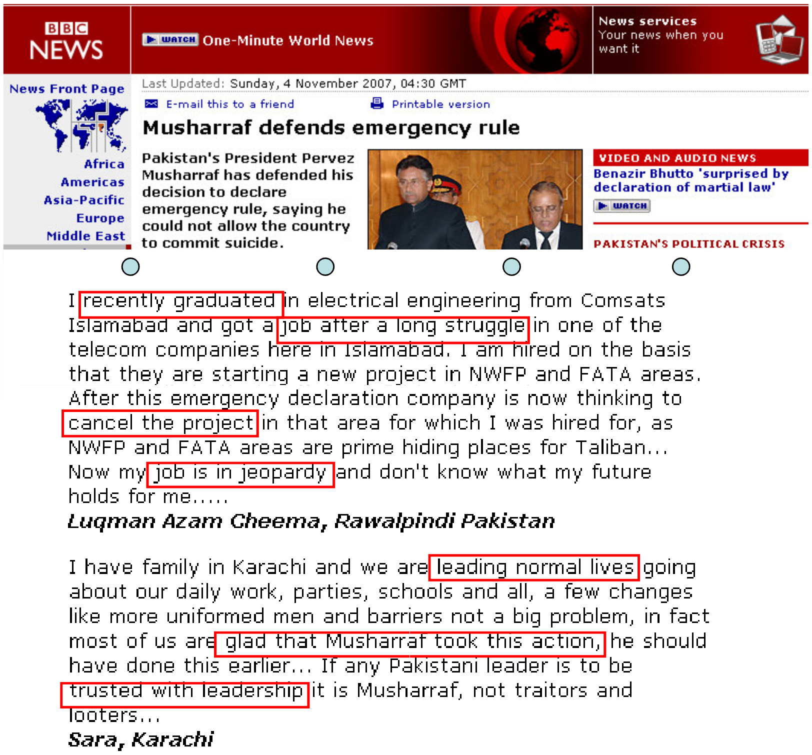
- Information that adds to the *completeness* of the picture about various matters.
- Information that *contextualizes* this complete picture by placing it in the right context for recipients to make it simpler and understandable to them.

As mentioned in Chapter 1, the usefulness for a user of a particular message will depend upon the type of information provided by it. Therefore, if we can model participatory media to be able to predict what type of information a particular message is likely to provide to a user, and if we can learn on a personalized basis the preferences of the user towards different types of information, we can make a better choice of which messages will be the most useful and should be recommended to the user. We will show in this chapter that social networks help us make this prediction, and in Section 5.5 we will eventually give an overview of how we use social network based modeling to design a personalized recommendation algorithm. But before that, we give a few examples to understand participatory media more closely.

Example 1: BBC News – Musharraf defends emergency rule

An article titled *Musharraf defends emergency rule* was published on November 4^{th} 2007 about the Emergency declared in Pakistan. The article described some aspects of the event, such as President Musharraf's justification of his decision, condemnation by other political leaders of the country, and reaction of the judiciary [1]. Two comments on the same article are shown in Fig.5.1.

[1] http://news.bbc.co.uk/2/hi/south_asia/7077310.stm



Last Updated: Sunday, 4 November 2007, 04:30 GMT

Musharraf defends emergency rule

Pakistan's President Pervez Musharraf has defended his decision to declare emergency rule, saying he could not allow the country to commit suicide.

I recently graduated in electrical engineering from Comsats Islamabad and got a job after a long struggle in one of the telecom companies here in Islamabad. I am hired on the basis that they are starting a new project in NWFP and FATA areas. After this emergency declaration company is now thinking to cancel the project in that area for which I was hired for, as NWFP and FATA areas are prime hiding places for Taliban... Now my job is in jeopardy and don't know what my future holds for me.....

Luqman Azam Cheema, Rawalpindi Pakistan

I have family in Karachi and we are leading normal lives going about our daily work, parties, schools and all, a few changes like more uniformed men and barriers not a big problem, in fact most of us are glad that Musharraf took this action, he should have done this earlier... If any Pakistani leader is to be trusted with leadership it is Musharraf, not traitors and looters...

Sara, Karachi

Figure 5.1: BBC News: Musharraf defends emergency rule

Consider a hypothetical recipient user to determine what type of information each of the above comments would provide to the user. If the recipient user is in fact in similar circumstances as the first message author, maybe a recent graduate or even a classmate of the author, then the first comment would provide useful information to the recipient because it would contextualize the article to the recipient. However, irrespective of the context of the recipient, the comment would also provide information that adds to the diversity of aspects about the event to convey a more complete picture to the recipient. Similarly, the second comment would provide completeness information to most recipients by adding a different opinion to the event, but provide simplification and understanding generally to only those recipients whose context is similar to that of the author. Therefore, both the comments would provide different types of information to different recipients, but if we can infer the contexts of the message authors and recipients, we might statistically be able to make good predictions about the type of usefulness of information provided to the recipients.

Example 2: The Economist – Malaria and how to beat it

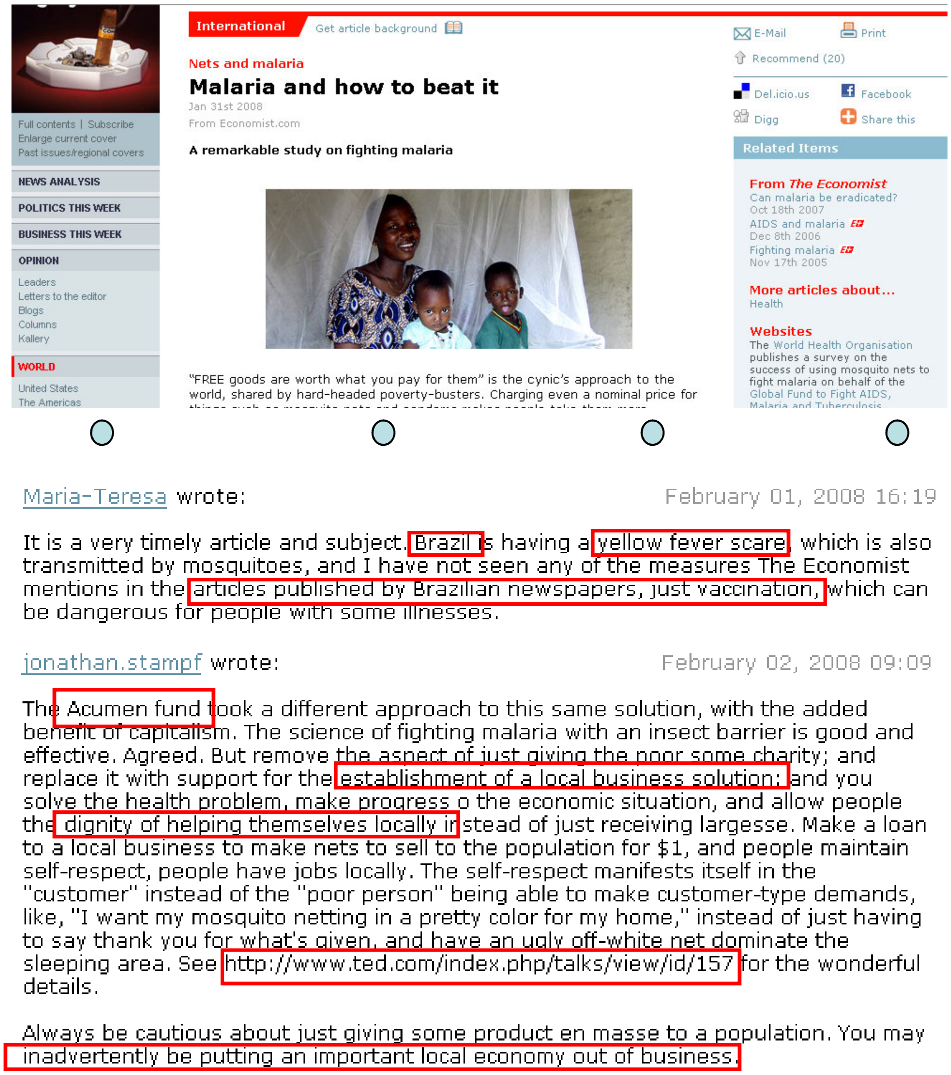
The Economist published an article titled *Malaria and how to beat it* on January 31st 2008, about a study in Kenya which concluded that malaria nets distributed for free produced better results than when they were sold for nominal prices [2]. The study was meant to counter the popular notion that people do not attach significant importance to goods unless they pay for the goods. Consider two comments on the article shown in Fig.5.2.

As in the first example, both these comments explored aspects of the topic that had not been examined in the original article and added to its completeness. However, the first comment would be particularly useful for Brazilian newspapers and healthcare workers because it would explain the relevance of the article to them. Similarly, the second comment would be quite useful for health agencies and policy makers because it would situate the article in the right context for them.

More examples are given in [191]. To summarize, the examples show that participatory messages can be useful to different recipients for different reasons depending on the context of the authors and recipients. In this thesis, we focus on two of these reasons to model the usefulness of a message for a recipient: given a message about some event, the message would be useful because of (i) the completeness of aspects about the event provided by the message, and (ii) the contextualization provided by the message to help the recipient understand the various aspects about the event.

The rest of the chapter is organized as follows. We define more precisely in Section 5.1 the concepts of context and completeness to be used in our framework for determining the usefulness of a message to a user. We then explain in Section 5.2 how social network information can help to infer the contexts of authors

[2] http://www.economist.com/daily/news/displaystory.cfm?story_id=10610398



Nets and malaria

Malaria and how to beat it

Jan 31st 2008
From Economist.com

A remarkable study on fighting malaria

"FREE goods are worth what you pay for them" is the cynic's approach to the world, shared by hard-headed poverty-busters. Charging even a nominal price for

Related Items

From *The Economist*
Can malaria be eradicated?
Oct 18th 2007
AIDS and malaria
Dec 8th 2006
Fighting malaria
Nov 17th 2005

More articles about...
Health

Websites
The World Health Organisation publishes a survey on the success of using mosquito nets to fight malaria on behalf of the Global Fund to Fight AIDS, Malaria and Tuberculosis.

Maria-Teresa wrote: February 01, 2008 16:19

It is a very timely article and subject. Brazil is having a yellow fever scare, which is also transmitted by mosquitoes, and I have not seen any of the measures The Economist mentions in the articles published by Brazilian newspapers, just vaccination, which can be dangerous for people with some illnesses.

jonathan.stampf wrote: February 02, 2008 09:09

The Acumen fund took a different approach to this same solution, with the added benefit of capitalism. The science of fighting malaria with an insect barrier is good and effective. Agreed. But remove the aspect of just giving the poor some charity; and replace it with support for the establishment of a local business solution; and you solve the health problem, make progress o the economic situation, and allow people the dignity of helping themselves locally instead of just receiving largesse. Make a loan to a local business to make nets to sell to the population for $1, and people maintain self-respect, people have jobs locally. The self-respect manifests itself in the "customer" instead of the "poor person" being able to make customer-type demands, like, "I want my mosquito netting in a pretty color for my home," instead of just having to say thank you for what's given, and have an ugly off-white net dominate the sleeping area. See http://www.ted.com/index.php/talks/view/id/157 for the wonderful details.

Always be cautious about just giving some product en masse to a population. You may inadvertently be putting an important local economy out of business.

Figure 5.2: The Economist: Malaria and how to beat it

and recipients, which can help to predict which messages will provide more completeness or contextualization to a recipient. This is followed by two sections of additional observations about the relationship between social networks and participatory media before we delve deeper into the recommendation algorithm. The first observation, presented in Section 5.3, is that of the dynamic nature of participatory media content as it evolves with more and more participation from users. The second observation, presented in Section 5.4, is the nature of credibility of the information given by participatory media content. Credibility of a message influences the usefulness of the completeness or contextualization information provided by the message. Finally, we put all these observations together in Section 5.5, and give an overview of our recommendation algorithm which learns the preferences of different users towards contextual, complete, and credible information in a personalized manner, and clarify how we use social network based modeling of participatory media content in our algorithm. Details of the algorithm and system design are then presented in subsequent chapters.

5.1 Context and completeness

Various terms have been in use to express the usefulness of a message for a user. In the field of information retrieval, a popular term is that of *relevance*, commonly used to describe the appropriateness of the topic relationship between a message and a query to retrieve the message [87]. However, this fails to take into account any specific requirements of the actual user making the query. Since different users could make the same query but have different requirements, or different perspectives necessary to understand the message, the term *pertinence* has been used to also consider the specific needs of a user [88]. Researchers have acknowledged though that usefulness, relevance, pertinence, etc are all multi-dimensional constructs [89], and characterized it through features such as the scope of the message and its understandability [90, 91]. Similarly, media researchers have explored the effects of information, and use terms such as simplification and opinion diversification to describe the effects a message may produce [12, 13]. We draw from these insights and define the following: Given messages about some event,

- **Context** *of* a message relates to its *understandability* with respect to a recipient [90], based on how well the message content explains the relationship of the message to the recipient. Thus, understandability can be considered as an outcome of the context of the message; messages that are more contextual will be more comprehensible to the recipient. Similarly, **context** *provided by* a message is related to the *simplification* of the event to a recipient [12]. Thus, simplification can also be considered as an outcome of the contextualization of the event provided by the message; messages that are more contextual will simplify the event for the recipient and explain its relevance to the recipient.

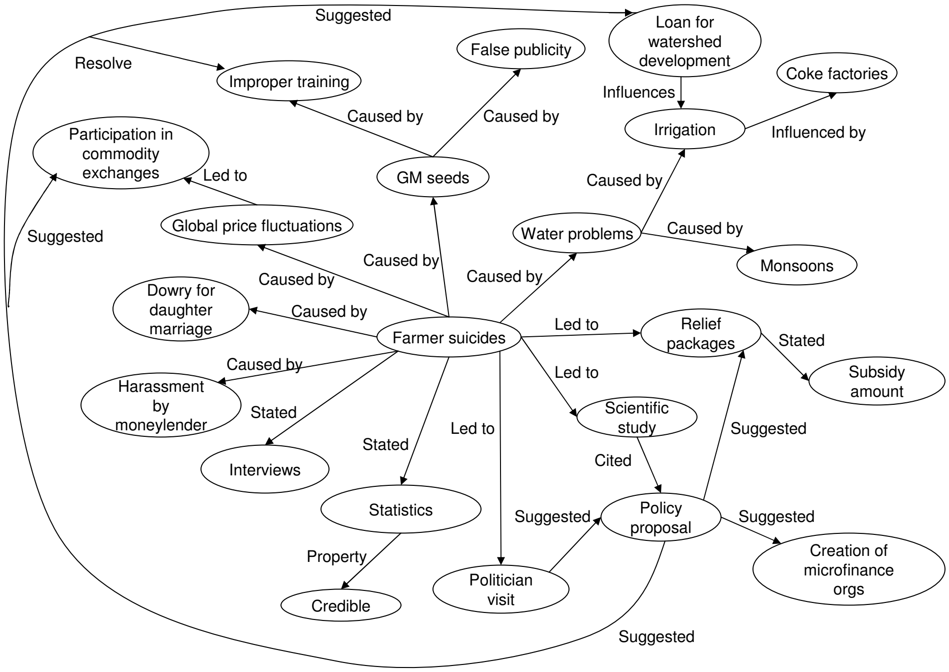


Figure 5.3: Sample ontology about farmer suicides in India

- **Completeness** *of* a message denotes the depth and breadth of aspects about the event covered by the message. A definition of *depth* and *breadth* is proposed by [92], as the depth and breadth of the topic ontology graph covered by the message. The *scope* of a message [90], or the *opinion diversity* expressed in the message [12], can be considered as an outcome of the amount of completeness of the message. Similarly, the additional **completeness** *provided by* a message denotes the additional scope or opinion diversity about the event provided by the message.

Note that context and completeness of messages are always observed from the perspective of a recipient (or *ego*), and are hence features of messages personalized for the recipient. Unless mentioned otherwise, context and completeness of messages will henceforth always be assumed to be stated with reference to some recipient. Referring to the examples given earlier, this terminology can be used to reiterate that the first comment in Example-1 provides *context* to classmates of the author of the comment, and the first comment in Example-2 provides *context* to health workers in Brazil. The comments also provide additional *completeness* to the readers by exploring aspects of the respective topics that had not been considered in the original articles. The following example is another attempt to make this more clear.

Example 3: Farmer suicides in India

Fig.5.3 shows an ontology of relationships between various aspects relevant to an event about suicides by cotton farmers in India [93]. Thousands of Indian farmers in the Vidarbha region of central India have committed suicide because of their inability to pay back loans they raised for cotton farming. This is attributed to many different reasons. For example, there was a sudden fall in global cotton prices that directly affected the farmers because the Indian government had to withdraw subsidies according to the WTO restrictions. In addition, the use of genetically modified seeds requires proper training for increase in yields, but training was not provided to the farmers. The inability to reuse GM seeds from the previous harvest further worsened the situation because farmers now had to purchase new seeds each time, without a proportionate increase in revenues. Other reasons include failure of the monsoon rains, lack of adequate irrigation facilities, and the corruption in getting loans from banks. It is important to identify the right reasons so that appropriate policies can be formulated to tackle the problem.

In this case, the completeness of a message about farmer suicides (the *event*) can be stated as the fraction of the aspects of the ontology graph covered by the message. Context provided by a message is harder to show in the figure, but it can be explained as follows by considering different recipients. For a moneylender, a message that explains how the creation of microfinance organizations in villages will impact his business, will provide context about the event to the moneylender. Similarly, for a politician, a message that explains how the provision of relief schemes in villages will impact the voting patterns in the next election, will provide context about the event to the politician. We show in the next section that knowledge of the social network of authors and recipients can help predict which messages will provide context and completeness to a recipient.

5.2 Role of social networks

Notice that message authors and recipients are embedded in an underlying social network of friendships and acquaintances. We next use insights from the *strength-of-weak-ties* hypothesis in social network theory [94] to develop a model that explains how context and completeness provided by messages may arise based on the implicit relationships between authors and recipients.

The *strength-of-weak-ties* hypothesis [94] states that social networks consist of clusters of people with “strong” ties among members of each cluster, and “weak” ties linking people across clusters. This is shown in Fig.5.4, assuming reciprocity in ties. Whereas strong ties typically link together close friends, weak ties link together acquaintances. The hypothesis claims that weak ties are useful for the diffusion of information, influence, and economic mobility, because weak ties help connect diverse clusters of people with each other, whereas strong ties may not bring about as much diversity. Implicit to the hypothesis is the phenomenon of

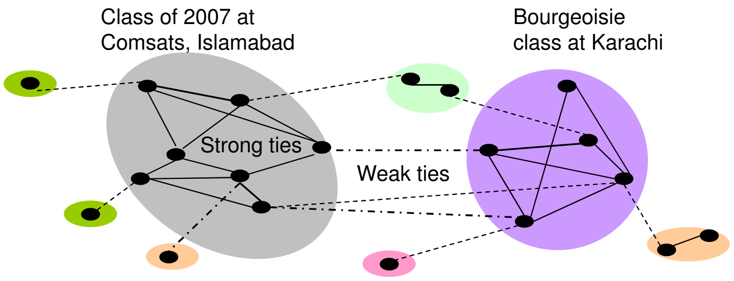


Figure 5.4: Strong and Weak Links

homophily noticed in multiple studies, that people similar to one another in terms of age, ethnicity, geographical location, income status, etc, tend to cluster together and reduce diversity [95, 96].

This insight can be applied to participatory media as follows. Messages about some event written by users strongly tied to a recipient are likely to be more simple and understandable, because these authors would have expressed similar perspectives as the recipient in interpreting the event. Thus, in the case of Example-1, the first comment would provide context to other recent college graduates, some of whom by the *homophily* argument are likely to be classmates of the author, connected to him by strong ties. On the other hand, the second comment would provide completeness, and by the *strength-of-weak-ties* argument, the author is likely to not be a strong tie of the college graduates. This is shown in Fig.5.4. Therefore, consideration of the relative position of the message author with respect to a message recipient in the social network, can predict the type of information that a message is likely to provide.

Many other studies have also made similar observations. [97] introduced the concept of *complex knowledge* as knowledge requiring more codification to be understandable, and showed that strong ties help understand complex messages, but weak ties are more useful to search for messages. [98] traced the changes in political opinion of people before and after the 1996 presidential elections in USA, observed with respect to the social networks of people. It was shown that weak ties were primarily responsible for the diffusion of divergent political opinion into localized clusters of people having strong ties between themselves. As indicated by the *strength-of-weak-ties* hypothesis, this reflects that local community clusters of people are often homogeneous in opinion, and these opinions may be different from those of people belonging to other clusters. We therefore hypothesize the following:

Messages written by users strongly connected to a recipient will provide more context than messages written by users weakly connected to the recipient, and messages written by users weakly connected will provide more completeness than messages written by users strongly connected to the recipient.

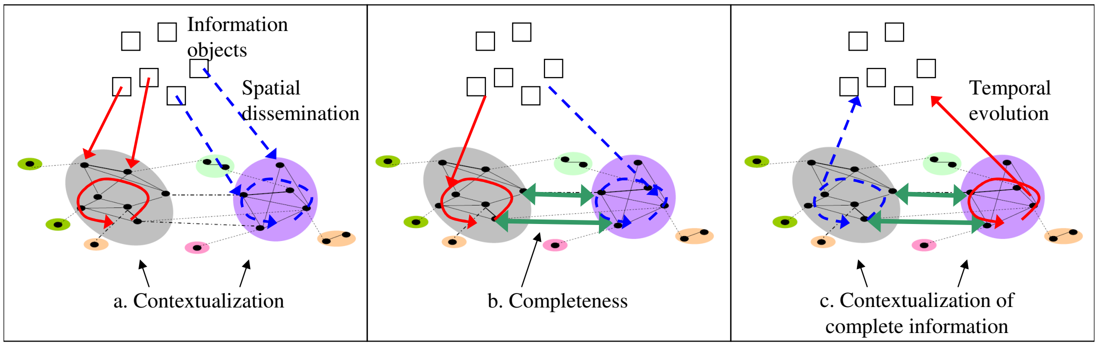


Figure 5.5: Ecosystem of information producers and consumers

A more detailed and precise form of the hypothesis [3] is validated in Chapter 6. This hypothesis will form our basis for using social network information to infer contexts and improve recommendations. We next describe two additional observations about the relationship between social networks and participatory media, and finally give an overview of the recommendation algorithm in Section 5.5.

5.3 Information evolution through participation

Implicit to this discussion is a dynamism at which we only hinted in Chapter 1 – that of information evolution. Information about an event evolves over time as users write messages about it, and other users write comments or blogs about these messages. Considering a particular recipient user, when messages gather contributions from users in the same strongly connected cluster as the recipient, the messages are likely to gain context. When the messages circulate among users in adjacent clusters connected through weak ties to the recipient's cluster, the messages are likely to gain completeness. Thus, messages can be visualized as *spreading* on the social network of users. Depending upon the pathways that the messages take relative to a recipient, they gain context and completeness. This process is shown in Fig.5.5 as operating in three stages.

- Part (a) shows that people in different parts of the social network may read different news articles, referred to as the *spatial dissemination* of news articles. Now, when people comment on the news articles locally within their clusters, it leads to *contextualization* of the news articles for other people in the same clusters.

[3] Our hypothesis is in fact valid only for *topic specific social networks* rather than the entire social network. Details are explained in the next chapter.

- However, when people read news articles and comments written by other people in adjacent clusters, it provides more *completeness* to them, as shown in Part (b). This can also happen when people cross-post across different clusters, and provide new viewpoints or address different aspects of the message topic.

- In part (c), the figure shows that these new viewpoints also circulate within external clusters to gain context. This process is termed as the *temporal evolution* of information, to indicate that the context and completeness of the news articles increases with time as they circulate among different participants. In practice though, all the above processes may occur concurrently and there may not be any time-sequencing between the three stages shown in the figure.

We believe that it is the unique capability of bi-directional one-to-many communication enabled through participatory media (Table 5.1) that creates a positive feedback system for information evolution by converting information consumers into information producers. The scenario can be easily generalized to information production and consumption across multiple clusters as well; we therefore consider participatory media as a communication process in an *ecosystem of information producers and consumers*. To summarize, Fig.5.5 aims to shows that if all entities in this ecosystem, irrespective of whether they are predominantly consumers of information or producers, will use recommendation algorithms that meet the goals we stated in Chapter 1, and if our hypothesis about the relationship of social networks to context and completeness given in the previous section is true, then such recommendation algorithms will enable the circulation of messages based on their context and completeness providing characteristics, and automatically lead to information evolution.

To demonstrate the value of information evolution enabled through participatory media, we refer to the example of the farmer as an information consumer given in Chapter 1, and generalize it to convert the farmer into an information producer as well.

Use case: The farmer as an information producer

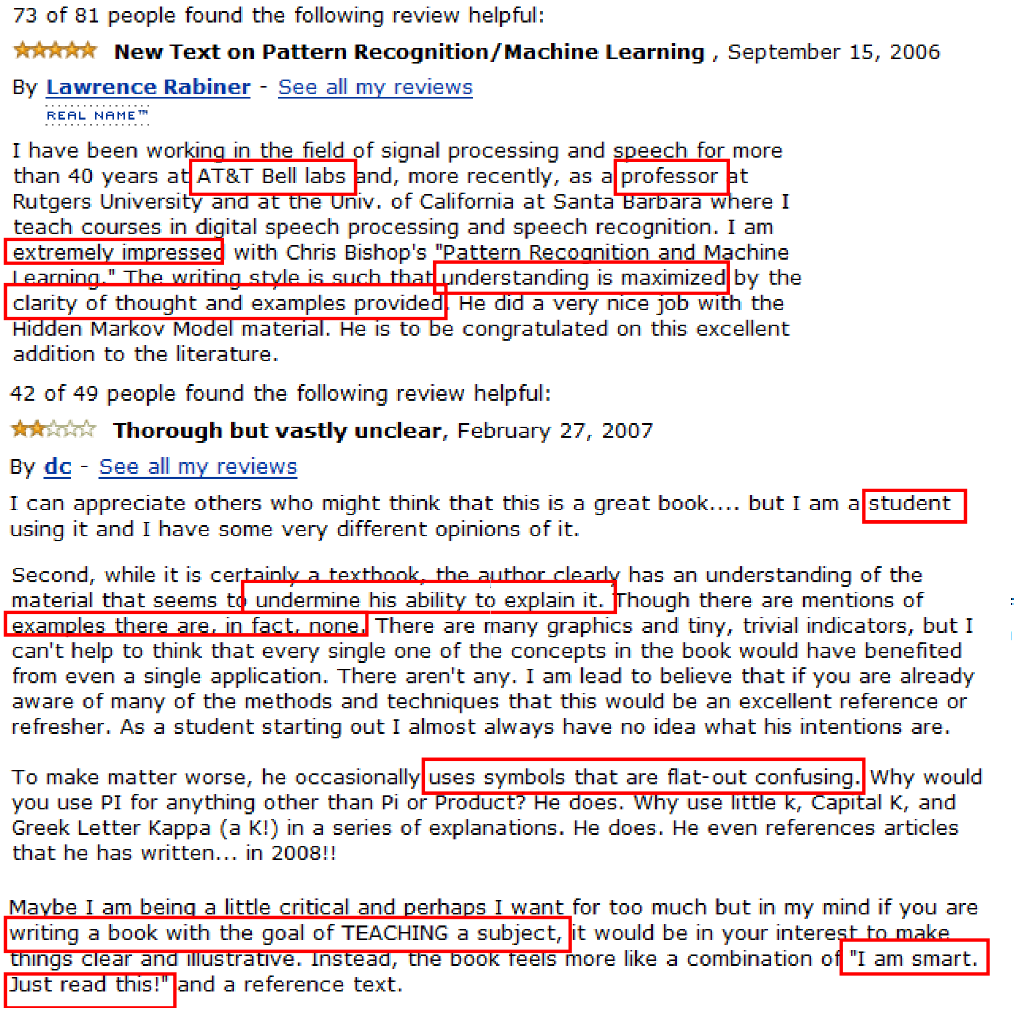
Being closely aware of the ground realities, the farmer is in fact the best source of information about the results of government policies for debt relief, status for the development of irrigation facilities, efficacy of the educational programs about best farming practices, etc. Feedback from farmers about this information can be useful for government agencies to track the success of various services provided by them. Considering governments, non-governmental organizations (NGOs), and politicians as additional users in this ecosystem of information consumers and producers, context and completeness based recommendation algorithms that we aim to develop can be expected to create desirable information circulation patterns as follows.

- *Part a*: Farmers would have strong ties with NGOs who work in their villages and know the local context [99]. Feedback about government schemes by the farmers will therefore be easily understandable by the NGOs.
- *Part b*: The NGOs would have weak ties with government agencies [100]. Therefore, aggregated farmer feedback given by the NGOs to the government can be expected to provide more diverse perspectives to the government about the situation of their schemes at the grassroots.
- *Part c*: Politicians would have strong ties with various government departments [101], and contextualization of the reports given by NGOs will help politicians understand the current deficiencies of government schemes, and propose changes for more efficient planning and implementation in the future.

Such information circulation patterns can create feedback and accountability loops for the provision of public services, and lead to the formulation of appropriate policies to tackle other economic problems that farmers may be facing [7]. We therefore conclude that if the unique features of participatory media that enable information evolution are supplemented with appropriate recommendation algorithms such as those described in this thesis, and increasing numbers of entities use such recommendation algorithms, the greater will be the benefits to society through appropriate information delivery. We would like to reiterate at this point that the nature of recommendation algorithms is of key importance to realize the benefits of human knowledge creation and its externalization into communicable information. Inappropriate algorithms can in fact prove detrimental to human society, for example, by causing users to waste their valuable time reading non-credible or non-contextual information, or by conveying a false sense of having received complete information. We will demonstrate the value of our particular algorithm for recommending messages to users, and reinforce the justification for a multi-disciplinary approach that integrates the fields of information science and media research to identify the factors of context, completeness, and credibility which define the appropriateness of recommendation algorithms.

5.4 Credibility

The assessment of credibility of messages is a crucial task for a recommendation algorithm because given the ease of publishing information on the Internet without any editorial checks by a formal agency, anybody can publish “incorrect” information, or bad-mouth “correct” information. However, judging the credibility of a message can be quite subjective. Consider the following example.



73 of 81 people found the following review helpful:

★★★★★ **New Text on Pattern Recognition/Machine Learning** , September 15, 2006

By **Lawrence Rabiner** - See all my reviews

I have been working in the field of signal processing and speech for more than 40 years at AT&T Bell labs and, more recently, as a professor at Rutgers University and at the Univ. of California at Santa Barbara where I teach courses in digital speech processing and speech recognition. I am extremely impressed with Chris Bishop's "Pattern Recognition and Machine Learning." The writing style is such that understanding is maximized by the clarity of thought and examples provided. He did a very nice job with the Hidden Markov Model material. He is to be congratulated on this excellent addition to the literature.

42 of 49 people found the following review helpful:

★★☆☆☆ **Thorough but vastly unclear**, February 27, 2007

By **dc** - See all my reviews

I can appreciate others who might think that this is a great book.... but I am a student using it and I have some very different opinions of it.

Second, while it is certainly a textbook, the author clearly has an understanding of the material that seems to undermine his ability to explain it. Though there are mentions of examples there are, in fact, none. There are many graphics and tiny, trivial indicators, but I can't help to think that every single one of the concepts in the book would have benefited from even a single application. There aren't any. I am lead to believe that if you are already aware of many of the methods and techniques that this would be an excellent reference or refresher. As a student starting out I almost always have no idea what his intentions are.

To make matter worse, he occasionally uses symbols that are flat-out confusing. Why would you use PI for anything other than Pi or Product? He does. Why use little k, Capital K, and Greek Letter Kappa (a K!) in a series of explanations. He does. He even references articles that he has written... in 2008!!

Maybe I am being a little critical and perhaps I want for too much but in my mind if you are writing a book with the goal of TEACHING a subject, it would be in your interest to make things clear and illustrative. Instead, the book feels more like a combination of "I am smart. Just read this!" and a reference text.

Figure 5.6: Amazon.com: Pattern recognition and machine learning

Example 4: Amazon.com – Pattern recognition and machine learning

Websites such as Amazon.com allow people to post book reviews. Consider the following reviews in Fig.5.6 given for a book titled *Pattern Recognition and Machine Learning (Information Science and Statistics)*, by *Christopher M. Bishop* [4].

Both the reviews appear to give contradictory opinions: which review should be considered more credible? The contradiction disappears if the context of the reviewers is considered. The first reviewer is a professor who has a good background

[4] http://www.amazon.com/Pattern-Recognition-Learning-Information-Statistics/dp/0387310738/

in statistics and machine learning, and it is quite possible that the examples given in book would have seemed sufficient to him. However, the second reviewer is a student who probably does not have a rich background in the subject, and hence found the book hard to read. This shows that credibility assessment of a review or comment is subjective, and it is unsurprising to find conflicting judgements. Note also that both the reviewers mentioned their role as a reviewer, that is, whether they were a professor or a student. This suggests that credibility judgement by people can also be influenced by the role of the reviewer because, for instance, people may believe a Professor's review to be more credible than that of a student.

Multi-dimensional construct

Such observations have been resolved by various researchers by modeling credibility as a multi-dimensional construct [102, 103]. For example, [102] reasoned about the credibility criteria used by people to judge the credibility of computerized devices and software, and identified the following different types of credibility:

- *Experienced*: This is based on *first-hand experience* of a user, and reflects her personal belief about a device or software.
- *Presumed*: This reflects personal biases of a user that give rise to *general assumptions* about certain categories of computing products; for example, presumptions based upon the company which developed the product, the cost of the product, the importance of the function performed by the product, etc.
- *Reputed*: This is based on *third-party reports* about different products.

Using the same approach, plus the insight that social networks help identify clusters of shared context, we distinguish between the following types of credibilities for messages:

- *Role based*: This is the credibility gained from the role of a message author as a professor or student or journalist, etc. The credibility associated with institutions such as the BBC, CNN, etc can also be considered as role based credibility.
- *Experienced credibility*: This is based on prior experience of a recipient with the message author, and reflects her personal belief about the author.
- *Public credibility*: This is based on the general public opinion about the author of the message, or about the message directly.
- *Cluster credibility*: Given the social network of the recipient in advance, cluster credibility is based on the opinion of other users in the same social network

cluster as the recipient. Since our hypothesis shows that users in a social network cluster are likely to share a similar context, therefore cluster credibility models the contextual nature of credibility which was shown in the Amazon.com example.

Different users are likely to have different propensities to rely on their own beliefs or on the public opinion or on other types of credibility. We will show later in Chapter 7 how the different types of credibilities can be mathematically quantified, and the propensities of individual users towards them can be learned in a personalized manner. We then consider the subjective credibility assessment of a message as a property of the contextual and completeness information provided by the message, and use it to adjust the usefulness estimate of the message for a recipient. In the next section, we explain how all of the models of context, completeness, and credibility can be tied together into a comprehensive recommendation algorithm.

5.5 Recommendation algorithm overview

Given a new message m_x, the goal of the **recommendation algorithm** is to produce an accurate prediction of whether a recipient user u will find m_x to be useful and credible. We use the following insights developed in the previous sections to design the recommendation algorithm:

- The context- and completeness-providing characteristics of messages can be inferred from the social network of the message author and recipients.

- The multi-dimensional credibility of messages can be inferred from knowledge of the social network of users and opinions about the messages expressed by other users.

- Users have different preferences for receiving contextual and complete information, and different propensities towards different types of credibilities.

A high-level overview of the algorithm along with pointers to chapters that contain a detailed description of various parts of the algorithm, is shown in Table 5.2. We begin with some assumptions such as having knowledge about the social network of the users, ratings given by the users, and topics of interest to the users. Note that these are reasonable assumptions to make in several real world environments. Knowledge about the social network and interests of users is becoming available through APIs provided by social networking websites such as Facebook (*www.facebook.com*). Information about message authors and ratings given by users is also made publicly available on most blogging websites (eg. *www.livejournal.com*) and knowledge sharing websites (eg. *www.digg.com*). Identity management of users across multiple websites is made easier through consortiums such as the OpenSocial initiative (*code.google.com/apis/opensocial/*) and other solutions for identity management. The algorithm uses this knowledge and insights gained from various fields of research to learn personalized models for users and produce message recommendations for them.

The algorithm meets the goals stated in Chapter 1. It makes personalized recommendations in a context-sensitive and complete manner. Credibility assessment is an inherent part of the algorithm. Scalability is also ensured, and the algorithm is strongly grounded in media theory to make its behavior explainable to both users and system designers. We are hopeful that the factors of context, completeness, and credibility are appropriate factors to ensure effective media delivery, and will help make better use of the unique features of participatory media, for example, by enabling efficient information evolution discussed in Section 5.3.

Table 5.2: Algorithm overview

1	*Assumption*: The social network structure is given in advance as an undirected graph $G(N, E)$, where each user is represented by a node $n \in N$, with edges $e \in E$ connecting n to any other users considered to be friends or acquaintances of n.	
2	*Assumption*: A list of topics T is given in advance, and a boolean value for each (user $n_i \in N$, topic $t \in T$) pair indicates whether or not the user is interested in the topic. The induced subgraph G_t formed by users who are interested in some topic t is called the *topic specific social network* for that topic.	
3	**Cluster** the topic specific social network graphs with the result that users within each cluster have strong ties between them, and users in different clusters are connected with weak ties.	Chapter 6
4	*Assumption*: Whenever an event e occurs related to topic t, many messages M about the event are created. Let A_M denote the authors of these messages, B_M their respective contributions, U_M usefulness ratings given to these messages by various users, and C_M credibility ratings given to the messages. The algorithm is assumed to have knowledge about A_M, B_M, U_M, and C_M for all messages $\in M$.	
5	For a recipient user n_i who has already read messages $M' \in M$ and given ratings $U_{i,M'}$ and $C_{i,M'}$ to these messages, learn two **Bayesian models** for the user: • *Usefulness model* : Learns n's preferences for receiving contextual and complete information: $\mathit{Usefulness}_{i,t}(U_{i,M'} \mid G_t,\ M',\ A_{i,M'}, B_{i,M'}, C_{i,M'})$. • *Credibility model*: Learns n's propensities for different types of credibilities: $\mathit{Credibility}_{i,t}(C_{i,M'} \mid G_t,\ M',\ A_{i,M'}, B_{i,M'})$ For a new message m_x, use these models to **infer** a prediction of whether n_i will find the message to be credible and useful. Combine the credibility and usefulness predictions to produce a final probability score to make a recommendation decision.	Section 7.1 Section 7.2 Section 7.3
6	*System design*: Model learning for each user and its use to make recommendation decisions can be a computationally intensive process; scalability should be ensured across hundreds of millions of users and tens of millions of new messages posted each day. A distributed system design is proposed to split the operations of the recommendation algorithm into centralized and distributed components, and improve the scalability of the system.	Chapter 8
7	The usefulness and credibility models aim to replicate the observed user behavior. However, the user may not have explored certain sections of the social network, and may not be aware of their usefulness. Although not mentioned in the goals stated in Chapter 1, the algorithm should be capable of *exploiting* the known user behavior to make accurate recommendations, while also *exploring* other parts to learn new user behavior. Management of the exploration-exploitation tradeoff has however been left to future work.	Chapter 9 (future work)

Chapter 6

Modeling of participatory messages

How would the worker come to face the product of his activity as a stranger, were it not that in the very act of production he was estranging himself from himself! – Karl Marx, Estranged Labour, 1844

In the previous chapter, we introduced the concepts of context and completeness provided by messages. The context provided by a message is related to the simplification of an event to a recipient, and the completeness provided by a message denotes the scope or opinion diversity about the event provided by the message. We then hypothesized about the relationship between the context- and completeness-providing characteristics of messages with the social network of message authors and recipients. Based on previous research in sociology, we reasoned that messages written by users strongly connected to a recipient will provide more context than messages written by users weakly connected to the recipient, and messages written by users weakly connected will provide more completeness than messages written by users strongly connected to the recipient. In this chapter, we state this hypothesis more precisely, and validate it through measurements and surveys of real users on a social networking website. We then develop a mathematical model to quantify the degree of context and completeness that is likely to be provided by messages given knowledge of the underlying social network of the authors and recipients; this model is used in Chapter 7 as an estimate for prior probabilities to learn the preferences of users towards receiving contextual and complete information. We begin by defining the notion of topic specific social networks, then present a precise hypothesis statement, followed by a quantitative validation of the hypothesis and model for context and completeness.

6.1 Topic specific social networks

Our hypothesis essentially states that clusters of strong ties of people define contextual boundaries between them, where people within the same cluster share a common context with each other. However, people have different interests, and clusters of common context should therefore be topic specific. Hence, we introduce the notion of a *topic specific social network*, which consists of only those people who are interested in a particular topic. Therefore, our hypothesis can now be restated as follows: topic specific social networks consist of clusters of people; messages about a certain topic written by people within the same cluster for that topic provide *context*, whereas more *completeness* is provided by messages written by people across clusters for the same topic.

6.2 Formal definitions

6.2.1 Context

We use the word *context* to denote a *set of circumstances considered in a communication task*, where the task may be the reading or writing of a message by a user. It is assumed that a message always refers to a real phenomenon, called an event. Therefore, for reading a message, context refers to the *circumstances considered by the reader for interpreting the event described by the message.* For writing a message, context similarly refers to the *circumstances considered by the author for describing the event.* Therefore, the following variable instances can be related together: a message m, an event e described in the message, a message recipient r, a message sender s, context considered by the recipient c_r, and context considered by the author c_s. Note that c_r and c_s are both instances of a set of circumstances. A few definitions and assumptions that form the basis of the hypothesis are presented next.

Definition – Contextual message: A message written in the same or similar context as that considered by the recipient, that is a message where $c_r \sim c_s$. Hence, it is also possible to consider $c_r \bigcap c_s$ as the degree to which a message is contextual to a recipient. Contextual messages are referred to as *context providing messages.*

Assumption – Hypothesis basis 1: Contextual messages provide simplification and understanding about the event to the recipient. Simplification of an event is thus assumed to be an outcome of reading contextual messages describing the event. If η denotes a measure of the degree to which a message m brings simplification to recipient r, then this assumption states that $\eta \propto |c_r \bigcap c_s|$.

Definition – Strong and weak ties: Granovetter proposed a linear combination of four characteristics for tie strength: the duration of the tie, the emotional intensity, the intimacy, and the reciprocity of the tie [94]. We use a slightly different definition for strong and weak ties. Given the undirected topic specific social

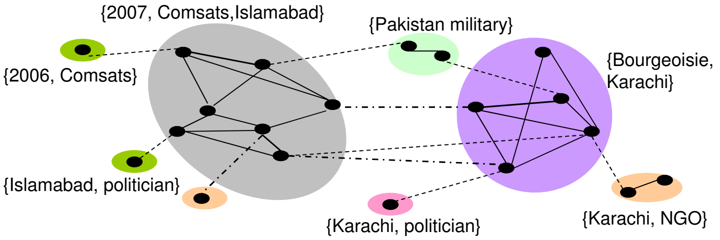


Figure 6.1: Context labeling of social network clusters

network of the topic to which e is related, we assume that we can use graph clustering algorithms to identify clusters of users such that nodes representing users within each cluster have a statistically significantly higher density of edges between than do edges to nodes representing users in adjacent clusters [104–106]. Then, we define the relationship between users in the same cluster as *strong* ties, and the relationship between users in neighboring clusters as *weak* ties. However, we make sure that our choice of clustering algorithm and its associated parameters result in a classification that agrees to a high degree with the notion of tie-strength as proposed by Granovetter. Note that considering rigid cluster boundaries is only a simplification for the model; in practice, such cluster boundaries are likely to be more diffuse [107]. We henceforth refer to the cluster of strong ties of the message recipient as b_r. Neighboring clusters to which weak ties exist from b_r are referred to as *adjacent* clusters b'_r.

Assumption – Hypothesis basis 2: Users in the same topic specific social network cluster share the same context for reading or writing a message about e. This can be conceptualized through Fig.6.1, which references Example-1 by hypothetically assigning context specifying labels to the social network clusters. Thus, the context considered by the author of the first comment could be specified as {*2007, Comsats, Islamabad*} whereas the context considered by the author of the second comment could be specified as {*Karachi, parties, bourgeoise*}. Our assumption therefore is that users in the same social network cluster share the same context for reading or writing messages. Note that even though we have indicated cluster labels in Fig.6.1, we do not make any claims about enumerating these labels; rather, our goal is to develop a notion of context similarity ($|c_r \bigcap c_s|$), which, as we will show, can be done without enumerating the underlying ontology of real world circumstances that give rise to various context labeling.

Hypothesis for context: It is now possible to combine the assumptions and definitions given above into a single testable hypothesis – for messages about an event e considered with respect to recipient r present in cluster b_r, messages written by authors $s \in b_r$ tend to be more contextual to r than messages written by authors $s' \in b'_r$. This leads to the first part of the hypothesis stated in Section 5.2:

Given a classification of ties as strong or weak in a topic specific social network, messages written by strong ties of a recipient provide more context than messages written by weak ties.

Experiments can now be designed to test this hypothesis in a straightforward manner. The first task is to identify a clustering algorithm that correctly produces a classification of ties into strong and weak. Then, the next step is to measure the outcome of the theoretical construct of context in terms of an observable factor which demonstrates the simplification or understanding η gained by recipient r from messages written by senders s and s'. The hypothesis will not be falsified if for a statistically significant fraction of recipients, $E_{s \in b_r}[\eta] > E_{s' \in b'_r}[\eta]$; that is, the mean simplification gained from messages written by senders in the same cluster as the recipients, is statistically greater than the mean simplification gained from messages written by senders in adjacent clusters. We will show in the next section how we survey users to estimate η and test for validation of the hypothesis. An alternative method is also proposed in Appendix B.

6.2.2 Completeness

We use the word *completeness* to denote the *degree to which relevant aspects of an event are analyzed*. Therefore, the following variables can be related together: an event e, a message m about the event, a set of aspects about the event discussed in the message a_m, and a universal set of aspects relevant to the event a_e. It is understood that $a_m \subseteq a_e$. Considering Example-3 about farmer suicides given in Section 5.1, the completeness of a message can be stated as the fraction of topics of the ontology graph covered by the message. Thus, a_e denotes the entire ontology graph about the event e, a_m denotes a subgraph covered by the message m, δ denotes the completeness of m, and $\delta \propto \dfrac{|a_m|}{|a_e|}$. This can be easily generalized into the mean completeness $E[\delta]$ of a set of messages M as well: $E[\delta] \propto \dfrac{|a_M = \bigcup a_m|}{|a_e|.|M|}$. The exact form of $|a_m|$ is left undefined; it can simply be the number of cells in Fig.5.3 covered by the message, or it can be a weighted sum based on the importance of different cells, or some other aggregate measure that even takes the nature of relationship between the cells into account. Similar to context, we seek a method that does not require an enumeration of the actual ontology to make our claims; the example is included here only for illustration.

Assumption – Hypothesis basis 3: The assumption is that users within the same cluster tend to focus on the same or similar aspects of an event. Hence, messages from adjacent clusters can be expected to provide more completeness. This can again be conceptualized through Fig.6.1, which shows various context labeling for different clusters. The assumption is that users within a cluster would tend to focus on some particular aspects only, and a more complete perspective can be formed by examining messages from adjacent clusters as well. Note that the assumption again does not require an enumeration of the underlying ontology of the real world; we are only concerned with the relative completeness of messages arising from different parts of the social network of users.

Hypothesis for completeness: The assumption and definitions given above can used to state a testable hypothesis for completeness – for messages about an event e, the mean completeness provided by messages written by authors $s \in$ some cluster b_r [1], tends to be less than the mean completeness provided by authors $s \in b_r \bigcup s' \in b'_r$. This leads to the second part of the hypothesis stated in Section 5.2, considering a cluster b_r of some specific recipient r:

Given a classification of ties as strong or weak in a topic specific social network, messages written by weak ties of a recipient provide more completeness than messages written by strong ties.

This hypothesis relies on the same algorithm mentioned earlier to find clusters of similar context. Experiments can now be designed to test the hypothesis in a straightforward manner, by measuring the outcome of the theoretical construct of completeness in terms of a measurable factor δ derived through **content analysis** or other means. The hypothesis will not be falsified if for a statistically significant fraction of clusters, $E_{s \in b_r}[\delta] < E_{s \in b_r \bigcup s' \in b'_r}[\delta]$; that is, the mean completeness of messages written by senders within the same cluster, is less than that of messages written by senders within the same and adjacent clusters.

6.2.3 Information evolution

We can use this formalism to also precisely state the information evolution dynamic that was described in Section 5.3. We define a temporal operator called *contextualization* for this purpose.

Contextualization: Given a message m about an event e, and a message m' produced in response to m, *contextualization* is denoted as $c \bullet a$, where a refers to the aspects about an event e considered in m, and c refers to the context considered in producing m'. Therefore, *contextualization* represents the process of producing a message in response to an earlier message to add context to the aspects of the event covered in the earlier message. Referring to the three steps shown in Fig.5.5:

- Part (a) shows that different messages about an event e are read by people in different clusters of the topic specific social network for that event. We refer to the context considered by users in the left cluster as c_1, and the context considered by users in the right cluster as c_2. Users within a cluster

[1] We use the same notation b_r to preserve consistency with the context formalization given earlier, but here b_r can be any arbitrary cluster independent of the recipient.

are strongly related to each other and consider the same context for reading messages: $c_{r1} = c_1$ and $c_{r2} = c_2$. These users add comments to the message or write related blogs, $c_{s1} = c_1$ and $c_{s2} = c_2$, and increase the context provided to other users in the same cluster. According to the context hypothesis, this helps users understand the event better. However, referring to the completeness hypothesis, the comments and blogs tend to focus on only certain aspects of the event, a_1 and a_2, for the left and right clusters respectively. Therefore, this process can be described by the *contextualization* operator as $c_1 \bullet a_1$ and $c_2 \bullet a_2$, to denote the contextualization of information about e within the clusters.

- Part (b) shows an exchange of messages or blogs across adjacent clusters along the connecting weak ties. This increases the completeness of information about e supplied to users in different contextual boundaries: $a_1 \bigcup a_2$ for both the clusters. However, unless this new information about e is not *contextualized* within these new boundaries, it does not help users understand the information and develop a more informed perspective about e.

- Part (c) shows that this information about e which flows across weak ties, is contextualized within the new boundaries. This is termed as the *temporal evolution* of information about e when messages about e circulate over social networks: $c_1 \bullet (a_2 \backslash a_1)$ and $c_2 \bullet (a_1 \backslash a_2)$. In other words, users in the left cluster may have ignored certain aspects of the event relevant to users in the right cluster $a_2 \backslash a_1$, believing that those aspects are not of significance to them. However, a contextual comment written by a user in the left cluster about these unconsidered aspects to the event, could help explain and convince the people that the information is actually relevant for them. It is therefore this process of temporal evolution of information that improves media effectiveness by providing people with an unbiased and better understanding of the event.

It is interesting to note that over time, this may even broaden the context considered on a regular basis by the people. This would especially be valuable in today's globalized world where local events can have widely dispersed global effects; very narrow contexts considered by people could turn out to be harmful for them, and society in general. Information circulation in this manner is likely to exhibit an emergent behavior with clusters of common context changing over time, but considering emergent behavior is beyond the scope of this thesis.

6.3 Quantitative modeling

In order to test our hypothesis about the relationship of social networks to context and completeness of messages, we crawled a popular social networking website, **Orkut**, and validated the hypothesis through surveys of real users. Based on the hypothesis, we then designed various graph theoretic measures as mathematical models for context and completeness, and validated the model in a similar way by correlating graph measurements with user-surveys. The quantitative model developed by us can be used to design personalized and context-sensitive information retrieval algorithms, such as the ones described in the next chapter. For formulating both the hypothesis and the model, we use the same notion of $|c_r \bigcap c_s|$ to infer the contextual similarity between a recipient and author, and $\frac{|a_M = \bigcup a_m|}{|a_e|.|M|}$ to infer the relative completeness provided to a recipient by messages written by different authors.

6.3.1 Preparation of the dataset

We wrote a web-crawler that screen-scraped a snow-ball sample from Orkut.com to obtain a social network graph of almost 800,000 users. Snowball sampling has a tendency to oversample hubs, and therefore we identified a core-set of approximately 42,000 users whose social network graph was known to a high degree of completion. This was done by recording only those users whose indegree was close to their outdegree in the initial larger dataset, making use of the evidence that the Orkut social graph has been seen to have a high degree of reciprocity [109]. This core-set of 42,000 users was used for further analysis, and the graph was considered as undirected for our experiments.

The graph followed a power-law in degree distribution with a truncation at 200, as also noticed in multiple other related studies [108, 109]. Orkut users can subscribe to various communities of interest and participate in discussions; we also crawled the community memberships for the set of users in our dataset, and a large number of discussions (ie. *messages*) in these communities. Orkut allows communities to link to other related communities; we then crawled the community graph and clustered it to derive coarse topics using a flow-stochastic graph clustering algorithm [110]. Some examples of topic clusters of communities that were identified were {*Books*, *Literature*, *Simply books*}, and {*Mumbai*, *Mumbai that I dream about*, *Mumbai bloggers*}. This indicates that related communities were indeed present in the same clusters to determine broad topics of interest. From this we were able to obtain the interests of users in different topics. Knowledge of user interests allowed us to extract *topic specific social networks* consisting of only those users and edges among users who were interested in a particular topic. We then selected four topic specific networks for our analysis: *Economics*, *Orissa* (a state in India), *Books*, and *Mumbai* (a city in India). Henceforth, any statistics about these clusters will be described in the same order.

6.3.2 Tie classification

Our hypothesis is based on the assumption that the nature of ties being strong or weak is given as a prior. We therefore first classify the ties between people in our dataset as being strong or weak, and then use this classification to validate the

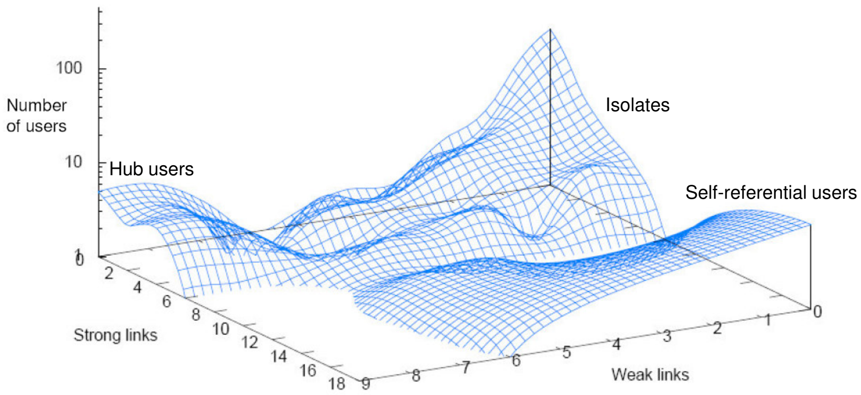


Figure 6.2: Mass-distribution of strong and weak ties in the topic cluster for Orissa

hypothesis. We assume that strong and weak ties can be differentiated from each other based on some clustering algorithm. Although significant research exists on the identification of such clusters [111], since we were agnostic to the actual choice of the clustering algorithm, we use the same flow-stochastic graph clustering algorithm [110] used earlier, to cluster the social network graph of all the 42,000 users. This algorithm has a configurable parameter to control the granularity of clustering, and hence produces different clusterings for different parameter values. We choose the parameter value that produces the closest agreement with user surveys, as described next.

We randomly chose {300, 250, 200, 500} users from the four topics respectively, and sent them a personalized survey in which we asked them questions about 5 of their randomly chosen **friends** with whom they had an explicitly declared reciprocal relationship. We asked these users to rate their 5 friends on a 5-point scale, giving a score of 1 to an acquaintance and a score of 5 to a close friend, and to also indicate the frequency of communication with them. Sample surveys are shown in Appendix C. This data allowed us to use communication frequency, emotional intensity, and reciprocity of a tie as proxies for the strength of a tie [94]. A total of 314 responses were obtained across the 4 topics, with ratings for 1,473 links. We then compared these ratings with the classification into strong and weak ties produced by the clustering algorithm.

The best choice of parameter gave a correlation of 0.76 between the classifications produced by the clustering algorithm and the classifications obtained from the user-surveys. The clustering produced by this choice of parameter was used for subsequent analysis. To give a visual explanation, Fig.6.2 shows the mass-distribution of the number of strong and weak links of users over the population of users interested in the topic of Orissa. It can be seen that a few *self-referential* users have many strong ties but few weak ties, while other *hub* users have many weak ties. The distribution characteristics in fact depend on the topic; other topic clusters have very different characteristics [191]. We will analyze these patterns across different topics in future work.

6.3.3 Hypothesis: Role of social ties

Given a classification of ties as strong or weak in a topic specific social network, messages by strong ties of a person provide more context than weak ties, and messages by weak ties provide more completeness than strong ties.

To test this hypothesis, we randomly chose 125 users per topic, and sent them a list of 5 of their friends who were also interested in the same topic. The users were not told which of their friends had been classified as strong or weak by our clustering algorithm. They were only asked to rank their friends on a 5-point scale to assess how much contextual and complete information each friend contributes to the user. We did this by framing a different question for each topic such that it captured the notion of context and completeness that we have defined. For example, we asked the users interested in Orissa to assume that they have to rely on their friends for the latest news about happenings in the state. Then we asked them to rank their friends based on how well the friends knew about their specific interests in Orissa ($\sim$ *context*), and how often the friends provided diverse viewpoints about happenings in Orissa ($\sim$ *completeness*). Note that via this method we attempt to directly estimate η and δ by asking the recipient users their opinion about the context- and completeness-providing characteristics respectively of their ties. We do recognize the shortcomings of our method as compared to an alternative method of using content analysis to estimate η and δ, but we feel our method to be sufficient as a first step. We have outlined an alternative validation experiment in Appendix B.

We received replies from {57, 46, 64, 63} users across the 4 topics, with information about {195, 204, 187, 188} links respectively. Each tie was then assigned three labels:

- {*strong, weak*}, given by the clustering algorithm.
- {*provides, does not provide*} context, given by the user surveys.
- {*provides, does not provide*} completeness, given by the user surveys.

Welch t-test: We produced two sub-samples of ties: (*strong tie*, {*provides, does not provide*} *context*), and (*weak tie*, {*provides, does not provide*} *context*). We then used the Welch t-test to compare the means of the first and second sub-samples by forming the null-hypothesis $\mu_1 = \mu_2$ and the alternative hypothesis $\mu_1 > \mu_2$ [112]. The mean of the first sub-sample was statistically much greater than the mean of the second sub-sample, and confirmed with a p-value < 0.001 (reject the null hypothesis) that strong ties are indeed more likely to provide context than weak

Table 6.1: Comparison of four scenarios: {*strong, weak* ties} promote {*context, completeness*}

	Context	**Completeness**
Strong ties	$\mu = .87, n = 133$ $z = 1.24^{***}$	$\mu = .50, n = 133$ $z = -0.38$
Weak ties	$\mu = .35, n = 71$ $z = -3.95$	$\mu = .70, n = 71$ $z = -0.88^{***}$

ties. In the same way, we produced two sub-samples of (*weak tie*, {*provides, does not provide*} *completeness*), and (*strong tie*, {*provides, does not provide*} *completeness*). Results again confirmed that weak ties are more likely to provide completeness than strong ties. This did not falsify our hypothesis about the relationship between social networks and the context and completeness of messages. We next proceed to analyze the data more closely, to study what proportion of strong and weak ties can be expected to provide context and completeness respectively.

Explicit scenario tests: We categorized our samples into four scenarios, {*strong, weak* ties} provide {*context, completeness*}. For each scenario, we performed the z-test by forming the null hypothesis (true mean $\mu = .8$) to indicate that at least 80% of the subjects believe in the scenario with an error-rate of 10% ($\alpha = 0.1$), and compared it with the alternative hypothesis $\mu < .8$ [112]. The choice of 0.8 as the true-mean is quite subjective, and only reflects an intuition that a majority of the subjects (aka. 80%) believe some scenario to be true. According to statistical tables, a z-value greater than -1.28 is considered as sufficient evidence to not reject the null-hypothesis. The results are shown in Table 6.1, and indicate that there is sufficient reason to not reject the claim that more than 80% of the subjects believe that strong ties provide context and weak ties provide completeness (marked as $***$). The test also succeeds for the scenario that strong ties provide completeness, although the mean is only 0.5, showing that strong ties also provide completeness but to a lesser extent than weak ties. Results from hypothesis tests on other topics are given in Appendix C.

6.3.4 Model formulation

We use the hypothesis validated in the previous section to develop a mathematical model to quantify the context and completeness of messages. We assume that we have knowledge of the topic specific social networks $G_t(N, E)$, connecting the users N with ties E. The clustering method outlined in the previous section is assumed to reveal tightly knit clusters in G_t, which are used to classify ties as strong or weak. We then assume that we also have knowledge of some messages M, authors of these messages, and users who have written comments on these messages. The modeling approach is as follows:

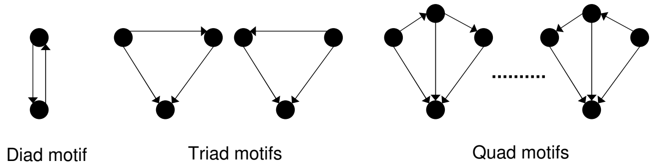


Figure 6.3: Motifs for clustering coefficients

1. The topic specific social network with clustering for topic t is derived as explained above.

2. The *active environment* of the message under consideration m_r is identified, defined as the set of participant users who have written a comment or reply to the message.

3. Wherever evident, we drop the subscript for topic t. For each cluster of strong ties V in topic t, its clustering coefficient C_V [113] is calculated. This is used as a proxy for the degree of contextual similarity among members of the cluster. We sometimes refer to C_{V_i} as the clustering coefficient of the cluster of user n_i.

- $\lambda_i = |\{\triangle' s \text{ centered on } n_i\}|$

  ie. define λ_i = number of triangles centered on user n_i, where $n_i \in N$. A triangle occurs when two neighbors of n_i are also connected to each other.

- $\tau_i = \binom{d^-}{2}$

  Here, d^- is the indegree of user n_i, and τ_i = maximum number of triangles that can be centered on user n_i.

- $c_i = \dfrac{\lambda_i}{\tau_i}$

  ie. define $c_i \in [0, 1]$ = clustering coefficient of user n_i.

- $C_V \in [0, 1] = \dfrac{\sum c_i}{|V|}$

  Here, $|V|$ denotes the size of the cluster. Thus, C_V is the average clustering coefficient for this cluster.

Fig.6.4 shows a scenario of four clusters A, B, C, and D, and the clustering coefficient computation.

The definition of the clustering coefficient proposed in [113] is clearly only one way of quantifying the cohesivity of a cluster [128]. Therefore, we experiment with three kinds of motifs, shown in Fig.6.3. The diad motif indicates the degree of reciprocity between ties, whereas the triad and quad motifs indicate the degree of cohesivity. Fig 6.5 shows the overall clustering coefficients of a few randomly

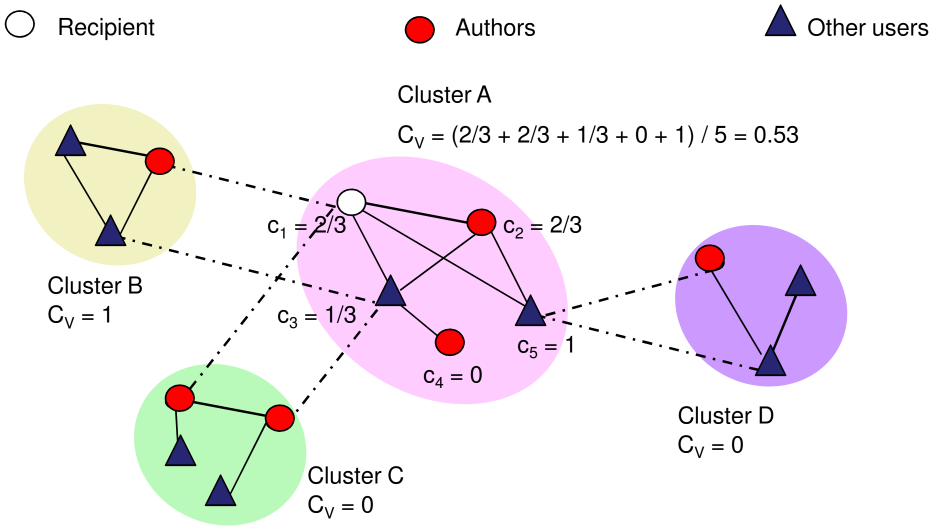


Figure 6.4: Computation of triad based clustering coefficient. Shown in the figure are four clusters: A, B, C, D. C_V^A, the clustering coefficient of cluster A, is calculated as the mean of c_i for all users $n_i \in A$. Here, $c_4 = 0$ because user n_4 has only one tie, and hence no pair of ties that form a $\triangle$. $c_5 = 1$ because user n_5 has one possible pair of ties within cluster A, and this pair does form a $\triangle$ with n_5. Proceeding in the same way, $c_3 = \frac{1}{3}$ because out of 3 possible pairs of ties, only one pair forms a $\triangle$ with user n_3.

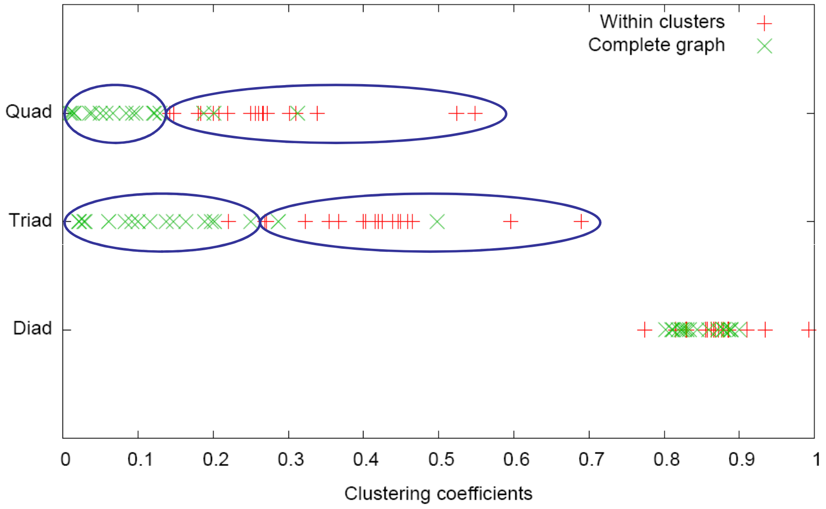


Figure 6.5: Clustering coefficients

selected topic specific networks, and the mean values of the clustering coefficients only within clusters of strong ties in each topic specific network. The diad-based clustering coefficient fails to differentiate within and across clusters because most ties in the Orkut network are reciprocal [109]. However, the triad- and quad-based clustering coefficients are quite successful; we therefore use both the triad and quad forms of the clustering coefficient for validation of the model.

4. For each participant user in message m_r, her integration coefficient into her cluster [114] is calculated. This is used as a proxy for the degree to which a user is embedded in her cluster, to capture the intuition that users deeply embedded in their clusters will provide more context.

$$\gamma_i = \frac{1}{(|V|-1)D_V} \sum_{n_j \in V} (D_V + 1 - d(i,j))$$

$d(i,j)$ is the distance from user n_i to n_j, calculated as the shortest path between the two users. D_V is the diameter of the cluster V = maximum distance between any two users $\in V$. Thus, the integration coefficient $\gamma_i \in [0,1)$ of user n_i into her cluster V, is close to 1 for users who are well integrated in their cluster, and close to 0 for users who are present along the boundaries of the cluster. Fig.6.6 shows the computation of the integration coefficient for users 2 and 4 in Cluster A. Closely integrated users have a high coefficient, while users present along the fringes have a lower coefficient.

Fig.6.7 shows the cumulative probability distribution of the integration coefficients of the users in the same randomly selected topic specific networks. The values are normally distributed, and the trend is consistent across different topics. We therefore consider the integration coefficient as a suitable metric to reflect the degree of embeddedness of a user into her cluster.

5. The context of message m_r for the recipient user n_i is calculated as follows:

$$Context_{ir} = \kappa C_{V_i} \sum_j \gamma_j$$

Here, the sum is taken over all users n_j in the active environment of the message who are also in n_i's cluster V_i. γ_j is the integration of user n_j into cluster V_i, and C_{V_i} is the clustering coefficient of the cluster of user n_i. The product of the clustering coefficient and integration coefficient is considered as a proxy for the amount of contextualization produced from a comment given by the participants, with the intuition that cohesive clusters with participants deeply embedded in the clusters will provide more contextualization. κ is a normalization constant to bring the value of $Context_{ir} \in [0,1]$, described later.

6. Let W_j^i denote those users from among the participants in m_r who are present in cluster V_j adjacent to the cluster of the recipient user n_i, and connected through weak ties to users in V_i. We define the second-degree integration of user n_i's cluster V_i into the active environment of m_r in cluster V_j, as follows:

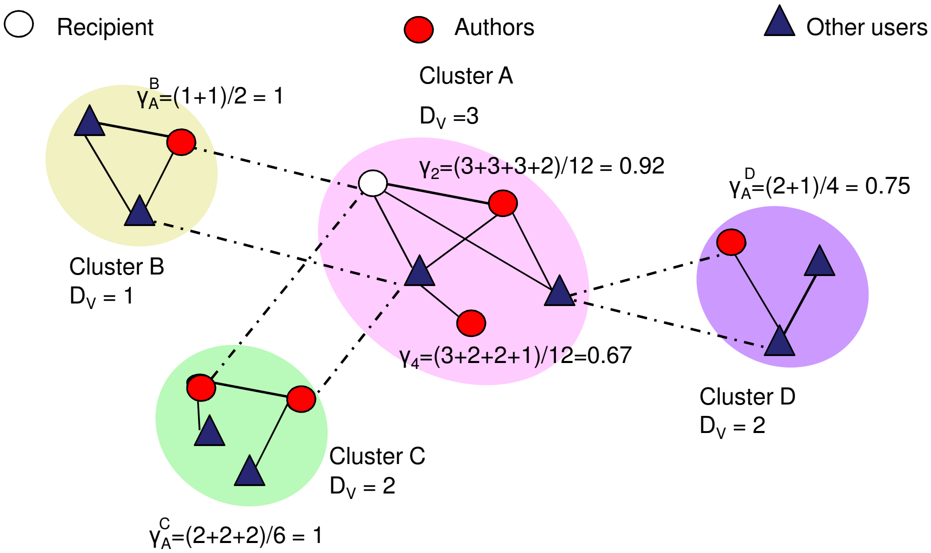


Figure 6.6: Computation of integration and second-degree integration coefficients. Shown in the figure is the calculation of the integration coefficients of message participants in cluster A, and the second-degree integration coefficients of clusters B, C, D into cluster A. Here, the integration coefficient of user n_4 is calculated by first enumerating the distances of n_4 to all other users in cluster $A = \{1,2,2,3\}$. Then, the diameter D_V of cluster A is found as the longest shortest path in the cluster, and is equal to 3. Next, the formula is applied which takes into account the sum of the difference between $(D_V + 1)$ and the distance of n_4 to other users $= \{3,2,2,1\}$. The greater the sum, greater is the integration coefficient. As can be seen from the figure, user n_2 is more closely integrated into the cluster than user n_4. Similarly, to give an example for calculation of the second-degree integration coefficient, we consider cluster D. Here, the participant is located at distances $\{1,2\}$ from other users in the cluster, and the diameter D_V of the cluster $= 2$. Hence, taking the mean of the difference between $(D_V + 1)$ and the distance of the participant to other users $= \{2,1\}$, gives a second-degree integration coefficient of 0.75.

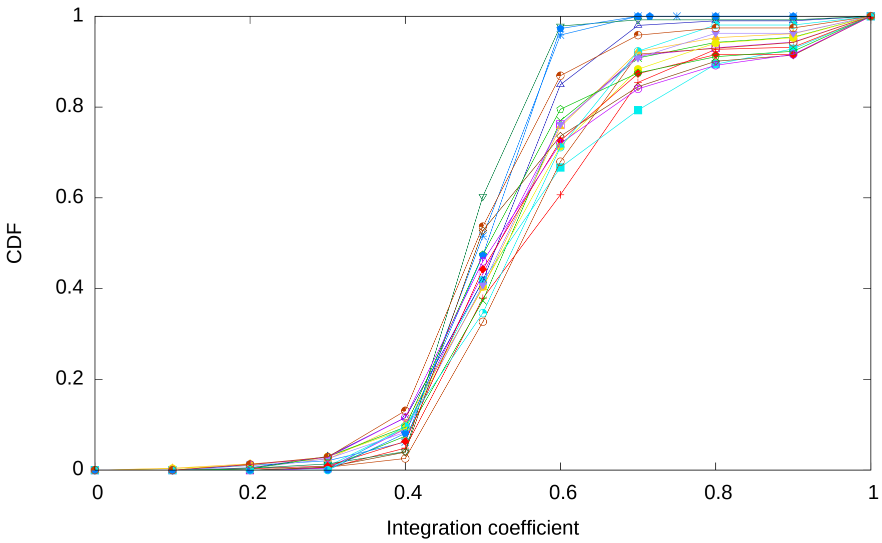


Figure 6.7: Integration coefficients

$$\gamma_i^j = \frac{1}{(|V_j| - 1)D_{V_j}} \sum_{n_k \in V_j} (D_{V_j} + 1 - d_j(W_i^j, k))$$

Here, $d_j(W_j^i, k)$ is the minimum distance to user n_k in cluster V_j from any user $\in W_j^i$. Thus, the second-degree integration of user n_i's cluster V_i into cluster V_j will be high if the weakly connected participants in a neighboring cluster are well distributed across the cluster, such that the minimum distance from a participant to every other user in the cluster is small. We will use the second-degree integration coefficient as a proxy for the degree to which participants in an adjacent cluster are embedded in that cluster, and are hence different from the cluster of the recipient user. Fig.6.6 shows the computation of the second-degree integration coefficients for clusters B, C, and D. Note that γ_{ii} can be calculated in the same manner, except that W_{ii} will include members from the same cluster as the recipient user.

7. The completeness of message m_r for user n_i is calculated as follows:

$$Completeness_{ir} = \kappa' \sum_j |V_j| \gamma_i^j$$

The sum is taken over all neighboring clusters V_j, where at least one user from V_j has contributed to message m_r. γ_j^i is the second-degree integration coefficient of cluster V_i into cluster V_j, calculated over only those weak neighbors of V_i who are present in V_j and the active environment of message m_r. Effectively, $Completeness_{ir}$ represents the area of the social network spanned by the message, reflecting the intuition that messages covering a greater part of the social network will be more complete. κ' is a normalization constant to bring the value of $Completeness_{ir} \in [0, 1]$, described later.

8. Since context and completeness are personalized measures with respect to the message recipients, we differentiate between the context and completeness provided

by the message to a particular user, and the average context and completeness provided by the message across all users. The average context of m_r is calculated as follows:

$$Context_{rt} = \kappa \frac{|V_i| C_{V_i} \sum_j \gamma_j}{\sum_i |V_i|}$$

Here, the mean is taken over all users who are interested in the topic t, and have at least one strong link from their cluster to some message participant.

9. The average completeness of m_r is calculated as follows:

$$Completeness_{rt} = \kappa' \frac{|V_i| \sum_j |V_j| \gamma_i^j}{\sum_i |V_i|}$$

The mean is taken over all users interested in the topic t, and have at least one weak link from their cluster to some message participant.

Analysis: Recall that our website crawl did not give us the social network of all users interested in a topic, but all topics in which our core-set of 42,000 users were interested. This data was suitable for testing the previous hypothesis because we assumed that post-pruning we had the complete real-world social network of these users. However, this data is not adequate to calculate the context and completeness of messages because it does not include the social network of all users participating in the message. Therefore, we selected 5 communities from each of the four topics, and crawled the list of all members of these communities. Then we crawled the complete list of friends of these members, so that eventually we had knowledge of the entire social network of users who were a part of these communities. We then randomly selected 118 discussions (ie. messages) from across these 20 communities, which gave us {44,28,17,29} messages from each topic respectively. For each message, we then framed a question that captured our notion of context and completeness as it would apply to a message recipient. For example, there was a discussion in a community for the development of Orissa, regarding good ways to use the Right to Information (RTI) law to identify places of corruption in government departments. We asked the selected users if the discussion outlined how they could use RTI in their specific circumstances (∼ *context*), and if the discussion covered other diverse circumstances for use of the law (∼ *completeness*). A 5-point scale was used for the ratings. We sent a survey for each message to 15 users, and received 837 replies. Examples of sample surveys are given in [191]. The ratings given in the replies were then compared with the measured values of context and completeness of the message for the users.

We first calculated the normalization constants κ and κ' as the right-hand intercept of the best-fit line (using linear regression) between the measured and surveyed

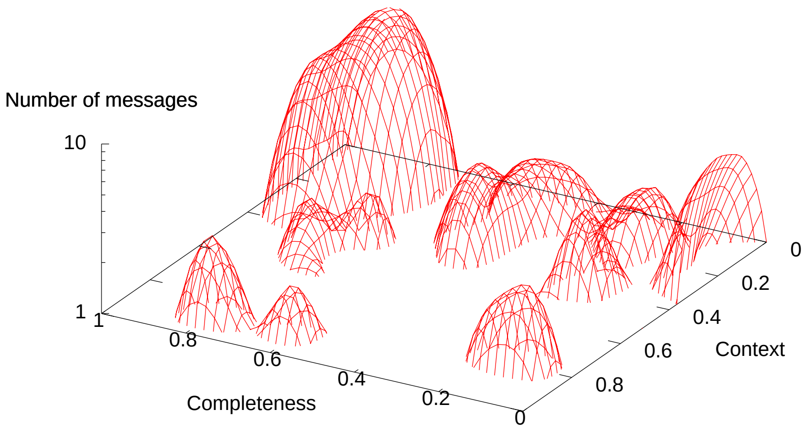


Figure 6.8: Mass-distribution of triad-based context and completeness of all messages

values for each message. The same constants were then used to normalize the measured values for average context and completeness of the message. Any values greater than 1 after the normalization, were capped at 1. This method has a limitation that it treats the ordinal Likert scale as an interval scale by assuming that survey values of (1,2), (2,3), etc have equal spacing across the scale. Therefore, we did the normalization for different spacing between the values, for example, by clumping the values at the edge or in the middle or assuming an exponentially varying spacing between the values. The tests described below gave similar results for all cases; we only present results of the equal spacing case.

Fig.6.8 shows a mass-distribution of the number of messages and their average triad-based context and completeness values over the population of selected messages. We divided the XY-plane into four quadrants for {*low, high* context} X {*low, high* completeness}. The quadrant boundaries were defined such that it results in approximately 10 messages per topic per quadrant. We then correlated these average measures of context and completeness with the weighted means of the context and completeness ratings given by the users. The same procedure was followed for quad-based measures of context.

The correlation coefficients (ρ_n for context, ρ_m for completeness) and t-test values (t_n for context, t_m for completeness) for triad-based clustering coefficients are shown in Table 6.2, when testing for the null hypothesis of the slope β of the regression line between the weighted means of the measured and survey values of context and completeness = 1 against the alternative hypothesis that $\beta < 1$. Although a few of the tests are not successful, the experiment does indicate that in most scenarios the measured values are able to give a good approximation to subjective user ratings (marked as $***$ for t-test values $\in [-1.64, 1.64]$). Results with quad-based clustering coefficients are in fact better, as shown in Table 6.3.

Table 6.2: Surveyed and measured values of message context (triad-based) and completeness

Context, Completeness	**n**	ρ	**t**
Low, low	n = 16	$\rho_n = .55$ $\rho_m = .79$	$t_n = -3.02^{*}$ $t_m = 0.15^{***}$
Low, high	n = 24	$\rho_n = .86$ $\rho_m = .92$	$t_n = 0.01^{***}$ $t_m = 1.09^{***}$
High, low	n = 36	$\rho_n = .81$ $\rho_m = .91$	$t_n = 1.11^{***}$ $t_m = 1.89^{**}$
High, high	n = 30	$\rho_n = .58$ $\rho_m = .90$	$t_n = -3.01^{*}$ $t_m = 1.09^{***}$

Table 6.3: Surveyed and measured values of message context (quad-based) and completeness

Context, Completeness	**n**	ρ	**t**
Low, low	n = 39	$\rho_n = .93$ $\rho_m = .76$	$t_n = -2.41^{*}$ $t_m = 0.19^{***}$
Low, high	n = 19	$\rho_n = .87$ $\rho_m = .84$	$t_n = 1.36^{***}$ $t_m = -1.71^{**}$
High, low	n = 26	$\rho_n = .86$ $\rho_m = .88$	$t_n = 1.43^{***}$ $t_m = 1.49^{***}$
High, high	n = 22	$\rho_n = .68$ $\rho_m = .88$	$t_n = -0.75^{***}$ $t_m = 0.04^{***}$

This gives further validity to the model, and shows that the model can be used to analyze the context and completeness of messages.

Note that the measures for context and completeness are time-dependent because the active environment of the message changes when more users contribute to the message [196]. Fig.6.9 shows the evolution of average triad-based context and completeness for a few messages from the Orissa community. The shaded area corresponds to the distribution of context and completeness in Fig.6.8 projected onto the XY plane. Each message corresponds to a line that shows a trace for context and completeness of the message, and each point on the line corresponds to a new user joining the active environment of the message. Thus, context and completeness of each message starts from (0, 0). Each time a new user contributes to the message, the context and completeness of the message changes, and the line advances to a new point.

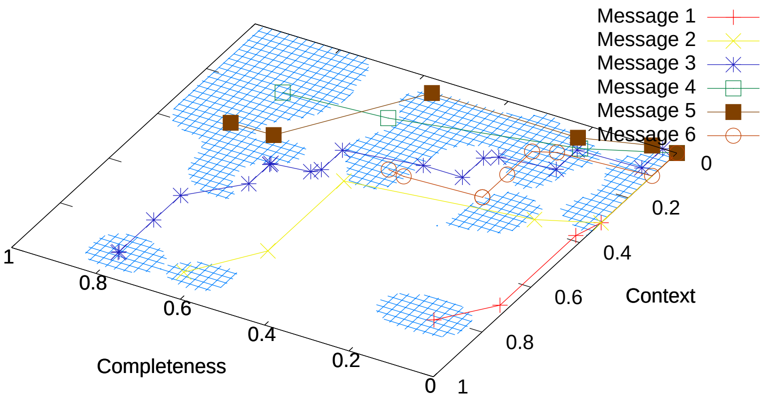


Figure 6.9: Evolution of context and completeness of messages

6.4 Discussion and related work

Our current approach for validating the hypothesis and model is limited by the brevity of the surveys used, and a single choice of clustering algorithm and metrics for context and completeness. We also recognize that since our evaluation has been done on only a single social networking website, it is improper to claim the validity of our hypothesis in all possible social network settings. However, the propositions made in this study clearly have much broader implications, and as discussed in Chapter 9, there is latitude to do further analysis of the theoretical constructs of context and completeness in future work.

The quantitative models for context and completeness can in fact be used in the following two ways:

- *User modeling*: As mentioned earlier, different users may have different personal preferences for receiving contextual and complete information. Machine learning techniques such as Bayesian networks can be used to learn personalized models for each user, which can serve as an extra filter for ranking search results or making recommendation decisions for messages. We describe such a technique in the next chapter.

- *Relevance ranking*: Document ranking metrics such as BM25F [115] provide a framework to incorporate customizable prior ranking probabilities along with TF-IDF scores and message lengths. These metrics can be extended to incorporate our mathematical models for context and completeness as well. We plan to experiment with this method in future work.

To the best of our knowledge, there is no prior research that has examined similar characteristics of context and completeness for messages. We therefore attempt to situate our work in this chapter with reference to other contemporary research activities on social networks. Most research can be grouped into the following three categories. First, there are purely measurement studies which have examined various graph-theoretic properties of different datasets. For example, [109] studied the link structure of users of four online social networking websites. Similarly, [116] studied the network of questions and answers in Usenet discussions to visualize *question-people* who asked questions, and *answer-people* who answered the questions. Second, there are studies which have applied insights gained from social networks to the design of applications. For example, [117] used social networks to improve web-page rankings produced by Google.Com. [118] inferred social networks in an e-commerce recommender system based on information flow patterns of transaction histories of users, and used the results to improve recommendation services. Third, there are studies which have proposed models for various scenarios in which social networks manifest themselves. For example, [108] proposed and evaluated a model showing that social network links among employees in a company tend to follow the lines of organizational hierarchy within the company. [96] proposed and evaluated a model showing that social network links created on the basis of geographical proximity can explain the small-world, navigability, and clustering properties of social networks.

Our work falls most closely in the third category, moving towards the second category. However, our work is distinguished from the third category because we explicitly focus on the manifestation of social networks on information characteristics rather than on other phenomenon such as geography [96] and organizational structure [108]. Our work is also different from that of the second category because we explicitly categorize messages in terms of their context- and completeness-providing characteristics, rather than treating messages as black-boxes as in [117,118]. In the next chapter, we show how to integrate message credibility into our metrics for context and completeness, and develop a comprehensive model for message ranking and recommendation to users.

Chapter 7

Personalized recommendation algorithms

There can be no liberty for a community which lacks the means by which to detect lies – Walter Lippman, 1920

In Chapter 5, we outlined the factors of context, completeness, and credibility which influence the usefulness of messages for users. Then, in Chapter 6, we proposed and validated a quantitative model to estimate the context- and completeness-providing characteristics of messages based on the underlying social network of the message authors and recipients. In this chapter, we use the model and insights about context, completeness, and credibility developed in earlier chapters to build a comprehensive **recommendation algorithm** for participatory media content. We experiment with two types of recommendation algorithms in this chapter, with the aim to accommodate the following requirements in both the algorithms:

- Ability to estimate the context providing characteristic of messages.
- Ability to estimate the completeness providing characteristic of messages.
- Ability to estimate the credibility of messages.

Our approach is to validate our algorithms by comparing their predictions against responses provided by real users as indicated in the **datasets**. As will become clear in later sections of this chapter, given the nature of the datasets, we are able to experiment with only a subset of the requirements for each type of algorithm. However, this does not imply that the algorithms cannot be generalized to accommodate all the three requirements given above. Preserving the terminology used in [120], the algorithms we develop can be classified as either **non-collaborative** or **collaborative** algorithms, and Fig.7.1 labels them in terms

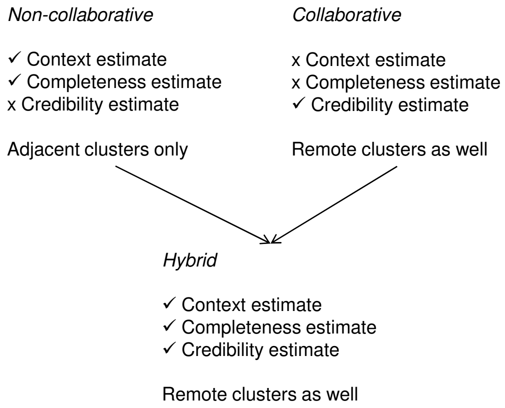


Figure 7.1: Recommendation algorithms

of the three requirements that we are able to evaluate for either algorithm. Section 7.1 describes the *non-collaborative* algorithm in which we show how to estimate the context- and completeness-providing characteristics of messages and learn the preferences of users towards these characteristics. Section 7.2 describes the *collaborative* algorithm in which we show how to estimate the credibility of messages, and learn the preferences of users towards different types of credibilities. Finally, Section 7.3 outlines a hybrid strategy for future work to combine the non-collaborative and collaborative approaches in a way to satisfy all the three requirements.

Also note that the context and completeness model we developed in Chapter 6, which forms the basis for the non-collaborative algorithm presented in Section 7.1, was formulated and validated only for adjacent clusters in the underlying social network. In fact, as shown in Section 6.3.4, a linear relationship was found to exist between the social network based measures for context and completeness, and the actual user perceptions of context and completeness of messages. However, it is quite possible that remote clusters which may not be exactly adjacent to each other may also share the same context. For example, with reference to Fig.6.1, politicians from Islamabad, Karachi, and the Pakistan military may share the same context, but their social network clusters are not immediately adjacent to each other. Therefore, a mechanism is needed to detect the contextual similarity between non-adjacent clusters, and to accommodate it in the models for context and completeness. The collaborative algorithm outlines such a mechanism which can be incorporated into the hybrid algorithm in the future. For now, we work with only adjacent clusters for the non-collaborative algorithm, but non-adjacent clusters for the collaborative algorithm.

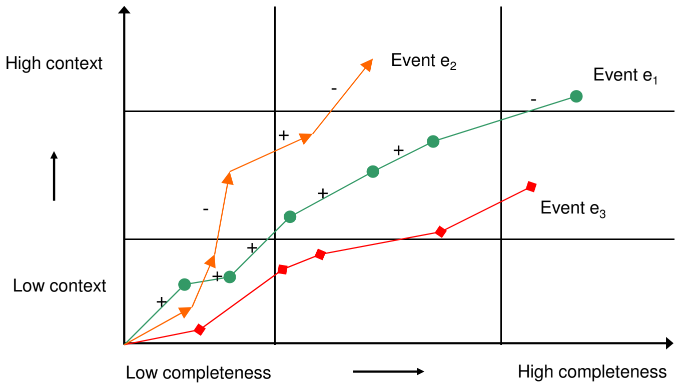


Figure 7.2: Knowledge state of a user

7.1 Non-collaborative algorithm

We refer to this algorithm as a non-collaborative algorithm because it only makes use of usefulness ratings given by a user to messages read by the user in the past, and uses these ratings to predict the usefulness of a new unseen message for the user. This is different from collaborative algorithms which make use of ratings given by other users to predict whether a particular user will find a new message to be useful. Non-collaborative algorithms therefore generally tend to be more privacy friendly than collaborative algorithms because patterns can be learned for each user independently. The difference between the non-collaborative algorithm described in this section, and the collaborative algorithm described in Section 7.2 will soon become clear, and eventually we will propose a hybrid algorithm which combines both these types of algorithms.

The non-collaborative algorithm works by **learning** a *usefulness model* for each user, which indicates the preferences of a user towards contextual or complete information. As stated in Section 5.5, the model can be learned given knowledge of the social network of the users, authors of messages seen in the past, and usefulness ratings given by a user to these messages. Once the model has been learned, it can be used to **infer** the usefulness of new messages that have not been seen by the user so far. We next explain the basis of the model, present some definitions, and then describe the model and our evaluation results in detail.

7.1.1 Basis of usefulness model

The usefulness model is based on the assumption that users have some inherent preferences towards the marginal utility they will gain from new message recommendations, where the utility gain will depend upon the additional amount of context or completeness provided by the message. This is shown schematically in Fig.7.2 to explain the intuition. The 2D space represents the *knowledge state* of a user with respect to some news event, quantified by the amount of contextual and complete information about the event the user has received at any point in time. A particular event or news story is represented as a broken line, and each kink on the line represents a new message about the story read by the user. For each story, whenever a new message is read, it changes the knowledge state of the user. This change will occur because the message will either provide more complete information, or help contextualize this information; and the change might be rated positively or negatively by the user (shown as ± link annotations in the figure).

Our assumption is that the message ratings given by the user on this 2D space will be consistent across stories because the user will have some inherent rating criteria based on the trajectory she prefers to follow in gaining contextual and complete information. For example, the user may prefer to gain complete information only if it is accompanied by messages that help contextualize this new information for her. Or, the user may prefer to follow a different trajectory of reading only more and more complete information. In addition, the user may prefer to read more and more information only to some extent, and the marginal utility she gains may become negative after a certain threshold is reached when a large amount of information has already been read by the user. Our goal therefore is to learn this preference function for each user over the 2D space of the current state of knowledge of the user. Once the function has been learned, it can be used to recommend messages to the user based on the prediction of how useful the user will find the messages to be.

The actual model we propose is more comprehensive. We take into account the topic of the news story, and learn a preference function of the user for each topic, rather than a single preference function for the user. We also take into account features such as the freshness and credibility of messages. Freshness is required to model traits of a user's behavior such as whether the user prefers timely recommendation of messages about some topic, and whether the user gains higher utility from the first message she receives about a story followed by lower utility from subsequent messages, etc. Credibility helps differentiate between messages supplying reliable versus unreliable information. Before describing the entire model, we first present some definitions.

7.1.2 Definitions

Message: A news article, or a blog entry, along with all its comments, is considered as a single message. The main component and all the comments in the message are individually referred as contributions to the message by users.

Message participants: This includes all users who have made contributions to the message.

Message environment: The underlying social network connecting the message participants is referred to as the message environment.

Message collection: A message collection is a set of similar messages, for example, a news story which is a collection of related news articles.

Knowledge state of a user: This represents all messages in a collection that have been read by the user. We quantify the current knowledge state of a user as the contextual and complete information the user has read so far.

Message freshness: This represents the timeliness of the message with respect to the period of relevance of the event to which it refers.

Message usefulness: Message usefulness is the rating given to a message by a recipient, for example, on a 5-point scale (1..5). In the proposed usefulness model, we assume that the usefulness rating given by a user is based on how much additional context and completeness is provided by the message, conditional on the current knowledge state of the user.

Broad topic: Each message or message collection will belong to a broad topic to which it is relevant. For example, a broad topic could be *books*, which will include messages such as book reviews, or prize announcements, etc. Similarly, *climate change* could represent another broad topic. We are presently uncertain as to what criteria we should use to automatically infer a suitable granularity to classify topics as broad or narrow. We will explain later that we instead rely on the users to choose their own levels of granularity.

7.1.3 Usefulness model

Problem definition: The problem can now be precisely defined as being able to predict the usefulness of a new message not seen by the user so far, based on the current knowledge state of the user. We do this by learning the parameters of a **Bayesian network** for each broad topic that is of interest to a user.

Knowledge requirements: As stated in Section 5.5, we make a number of assumptions for having information about messages, message environments, etc, that are required to learn the usefulness model for a particular recipient user. These are as follows:

- All message participants who have been involved in the message so far.
- The message environment, that is, the topic specific social network of the recipient user and the message participants. Note that derivation of topic specific social networks requires knowledge of the overall social network graph and the topics of interest to users. The topic specific social network for a particular topic is then obtained as the induced subgraph of users interested in that topic.
- Messages from the same message collection (and the associated message participants and message environments) read by the recipient user in the past.

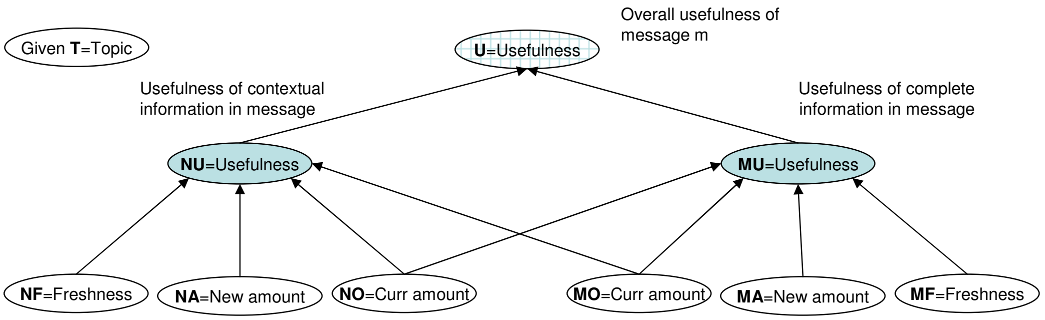


Figure 7.3: Usefulness model

- Archived data for usefulness ratings of messages from the same broad topic seen by the user in the past [1].

Note that in many real world contexts, this knowledge is in fact readily available. Knowledge about the social network and interests of users is becoming available through APIs provided by social networking websites such as Facebook (*www.facebook.com*). Information about message authors and ratings given by users is also made publicly available on most blogging websites (eg. *www.livejournal.com*) and knowledge sharing websites (eg. *www.digg.com*). Identity management of users across multiple websites is made easier through consortiums such as the OpenSocial initiative (*code.google.com/apis/opensocial/*) and other solutions for identity management. If we assume that the usefulness model will be implemented as part of a client-side application as described in the next chapter, then the application can even keep track of messages read (or clicked) by the user. We next describe the actual model, formulation for the evidence variables to learn the model, and finally present an evaluation of the algorithm.

We represent the usefulness model as a Bayesian dependency graph, as shown in Fig.7.3. Our decision to use Bayesian networks is motivated because of their support for causality [121], which help us directly model personalized features of users in terms of context and completeness. Here, directed edges indicate a dependency from the originating variable to the target variable. Shaded ovals (**NU**, **MU**) represent hidden variables and unshaded ovals represent evidence variables. The partially shaded oval for message usefulness (**U**) is a variable denoting the rating given by the user, and is available during the **training phase** only. The goal is to infer this variable for a new message, given the evidence variables and the parameters of the learned model.

The message usefulness **U** is assumed to depend upon the two hidden variables **NU** and **MU** for contextual and completeness usefulness respectively, provided by

[1] In general, learning and inference based on prior history always face a problem of cold start for new users. The standard method to solve this is to use randomization or recommendation of popular messages during the initial stages when sufficient data is not available [120]. However, we do not explore the cold-start problem in this thesis and leave it to future work.

the message. The hidden variables for contextual usefulness depend on evidence variables for the new amount of context provided by the message (**NA**), the current state of knowledge of the user (quantified as **NO** and **MO**, the current amounts of contextual and complete information respectively read by the user so far), and the freshness **NF** of the contextual component of the message. The dependency relationships for the completeness hidden variable are exactly similar. The message usefulness is expected to be positively correlated with **NA**, **MA**, **NF**, and **MF**, but negatively correlated with **NO** and **MO**. We next describe methods to quantify the evidence variables.

7.1.4 Calculating evidence variables

We use the model proposed in Chapter 6 to estimate the evidence variables for context and completeness provided by messages. Details of the formulation are repeated for readability.

Input

1. The social network structure is known in advance as an undirected graph $G(N, E)$, where users are represented as nodes N with edges E between users who declare themselves as friends of each other.

2. A list of broad topics T is known in advance, and a Boolean value for each (user n_i, topic $t \in T$) pair is also known that indicates whether or not the user is interested in the broad topic. The induced subgraphs formed by users who are interested in the same broad topics are referred to as topic specific social networks.

3. A graph clustering algorithm [110] is used to cluster topic specific social networks and classify ties between users within each cluster as strong ties, and ties between users in adjacent clusters as weak ties. Each cluster denotes contextual boundaries implying that users within each cluster share a common context with each other. Note that a shortcoming of this representation is that it restricts a user to be a member of only one cluster. In the future, we will extend the representation to allow users to be members of multiple clusters as well.

Contextual variables: *no* and *na*

4. We consider the evidence variables for context provided by messages as being composed to three components: a social network based estimate of the context provided by the message, the length of the message, and the credibility of the message.

4a. **Social network based component**

4a.i. For each cluster of strong ties V, its triad-based clustering coefficient C_V is calculated [113]. The clustering coefficient is used as a proxy for the cohesivity of the cluster, to denote the degree of shared context among members of the cluster. We sometimes refer to C_{V_i} as the clustering coefficient of the cluster to which user n_i belongs. Note that these calculations are done separately within each topic specific social network.

4a.ii. For each user n_i, her integration coefficient γ_i into her cluster is calculated [114]. The integration coefficient is used as a proxy for the amount of contextualization provided by messages written by the user.

$$\gamma_i = \frac{1}{(|V|-1)D_V} \sum_{n_j \in V} (D_V + 1 - d(i,j)) \tag{1}$$

Here, $d(i,j)$ is the distance from user n_i to n_j, calculated as the shortest path between the two users. D_V is the diameter of the cluster V = maximum distance between any two users $\in V$. Thus, the integration coefficient $\gamma_i \in [0,1)$ of user n_i into her cluster V, is close to 1 if she is well integrated into her cluster, ie. they are close to many other users. Similarly, γ_i is close to 0 if she is present along the boundaries of the cluster and is not well integrated.

4a.iii. We assume that for each event and its corresponding message collection M, the set of messages $m_j \in M$ seen by the user in the past is known. The message participants of the j^{th} message are denoted as $A(m_j)$, and their individual contributions are denoted as $B(m_j)$. For participants linked through strong or weak ties with a message recipient, the participants are denoted as $A_s(m_j)$ or $A_w(m_j)$ respectively. The same convention is followed for their respective contributions: $B_s(m_j)$ or $B_w(m_j)$. m_x denotes a new message.

4a.iv. The social network based component of the context provided to user n_i by the j^{th} contribution of message m_x belonging to a broad topic t, is denoted as $SNContext_{ijxt}$. We skip the subscript t for simplicity, and model the social network based component as follows:

$$SNContext_{ijx} = C_{V_i}.\gamma_j \tag{2}$$

Here, C_{V_i} is the clustering coefficient of the cluster of the message recipient n_i. γ_j is the integration coefficient of the j^{th} message participant (eqn. 1). Note that this is valid only for message participants in the same cluster as the recipient user (γ_j is 0 otherwise). Thus, more context will be provided by a contribution from a user closely integrated into the cluster of the recipient user. Fig.7.4 and 7.5 show the computation of the social network based component of context for a message collection across two stages as the message acquires more contributions over time.

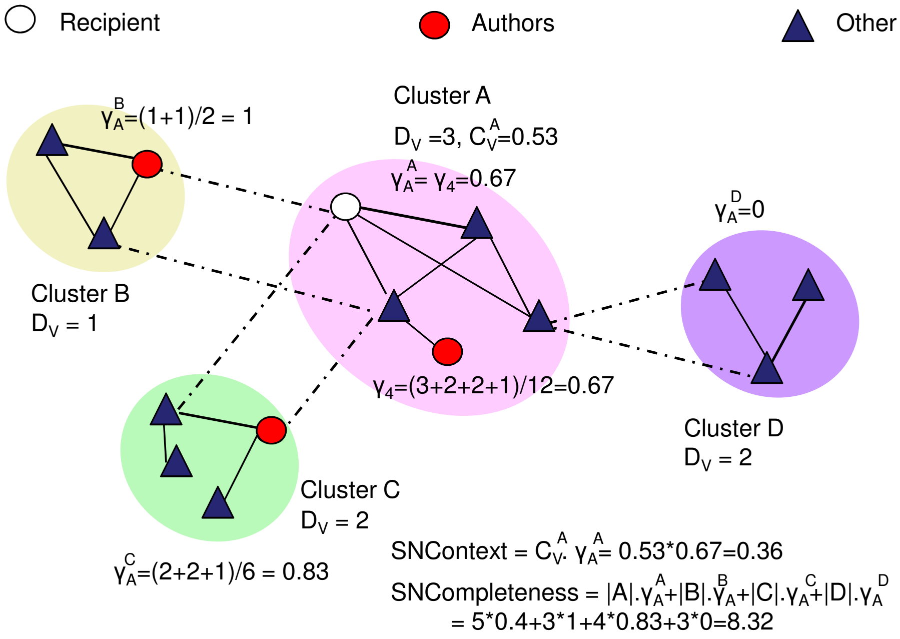

○ Recipient
● Authors
▲ Other users
γ_A^B=(1+1)/2 = 1
Cluster A
D_V =3, C_V^A=0.53
γ_A^A= γ_4=0.67
γ_A^D=0
Cluster B
D_V = 1
γ_4=(3+2+2+1)/12=0.67
Cluster D
D_V = 2
Cluster C
D_V = 2
γ_A^C=(2+2+1)/6 = 0.83
SNContext = C_V^A. γ_A^A= 0.53*0.67=0.36
SNCompleteness = |A|.γ_A^A+|B|.γ_A^B+|C|.γ_A^C+|D|.γ_A^D
= 5*0.4+3*1+4*0.83+3*0=8.32

Figure 7.4: Computation of context and completeness: stage 1

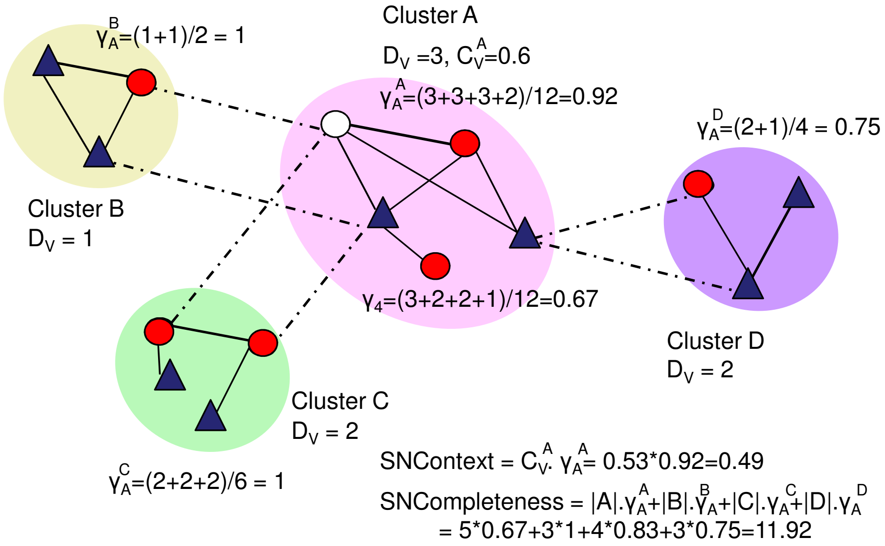

γ_A^B=(1+1)/2 = 1
Cluster A
D_V =3, C_V^A=0.6
γ_A^A=(3+3+3+2)/12=0.92
γ_A^D=(2+1)/4 = 0.75
Cluster B
D_V = 1
γ_4=(3+2+2+1)/12=0.67
Cluster D
D_V = 2
Cluster C
D_V = 2
γ_A^C=(2+2+2)/6 = 1
SNContext = C_V^A. γ_A^A= 0.53*0.92=0.49
SNCompleteness = |A|.γ_A^A+|B|.γ_A^B+|C|.γ_A^C+|D|.γ_A^D
= 5*0.67+3*1+4*0.83+3*0.75=11.92

Figure 7.5: Computation of context and completeness: stage 2

4b. **Message length**: The component based on message length is used as a proxy for the amount of information conveyed by the message. This simple heuristic can be extended by using language models and other information-retrieval techniques for length normalization [115]. In this thesis, we assume that a global constant L_t for topic t is known, such that the information content of the j^{th} contribution of message m_x is measured as:

$$info_{jx} = \min(\text{length}(b_j),\ L_t) \tag{3}$$

Here, b_j is a contribution made by the j^{th} participant in message m_x, considered only for message participants who are in the same cluster as the recipient user. L_t denotes a maximum threshold length for contributions.

4c. **Message credibility**: The credibility of each message participant is considered as a proxy for the credibility of the contribution made by the participant. In Section 7.2, we will show how to extend the credibility computation of messages by taking the ratings given by other users into account. However, in this section, the component for credibility can be simply expressed as:

$$Cred_{jx} = Cred'_j \tag{4}$$

Here, $Cred'_j$ denotes the credibility of the j^{th} participant for the topic to which the message belongs, and is naively calculated as the number of contributions made by the participant to the topic. As before, this is valid only for those users who are in the same cluster as the recipient user.

4d. NA = Evidence variable for the *new amount of context provided by a message*: Referring to Fig.7.3, we now need a method to combine the three components to calculate the contextual amount of information provided to user n_i by message m_x. However, there is no theoretical basis for combining these components except that they should be positively correlated with NA. We do believe that we need to identify a global definition for the function to calculate NA in a uniform way for all users, because NA estimates the contextual amount of information in any message. We represent this as follows:

$$(na)_{ix} = \sum_{j \in A_s(m_x), B_s(m_x)} f(SNContext_{ijx}, info_{jx}, Cred_{jx}) \tag{5}$$

In the evaluation, we experiment with different functions f to combine the three components in a product form, and choose the function that gives us the best performance. However, the actual function can be inferred through statistical learning when data on a large number of users is available, a subject we leave for future work. Finally, the sum of the contextual information provided by each contribution is then considered as the overall contextual information provided by the message.

4e. NO = Evidence variable for the *current amount of context already provided to the user*: Referring to Fig.7.3, this is essentially the sum of the context provided by individual messages from the same message collection that have been seen by the user, and is expressed as:

$$(no)_i = \sum_{m_x \in M} (na)_{ix} \tag{6}$$

Completeness variables: *mo* and *ma*

5. We similarly consider the completeness provided by messages as also being composed of three components: a social network based estimate of the completeness provided, the message lengths, and the message credibilities.

5a. **Social network based component**

5a.i. We first introduce the concept of a second-degree integration metric of user n_i's cluster V_i into an adjacent cluster V_j. Let W denote a subset of the destination nodes of weak ties from the cluster of user n_i into cluster V_j, where V_j is not the same as V_i. Calculate the second-degree integration as follows:

$$\gamma_{ij}(W) = \frac{1}{(|V_j| - 1)D_{V_j}} \sum_{n_k \in V_j} (D_{V_j}+1-d_j(W,k)) \tag{7}$$

Here, $d_j(W,k)$ is the minimum distance to user n_k in cluster V_j from any user $\in W$. Thus, γ_{ij} will be high if the subset of the destination nodes of weak ties into the neighboring cluster are well distributed across the cluster, such that the minimum distances from the nodes $\in W$ to every other user in V_j is small. We use the second-degree integration coefficient as a proxy for the degree of completeness provided by the adjacent cluster, to capture the intuition that a larger subset of weak ties into an adjacent cluster will provide more completeness. Note that $\gamma_{ii}(W)$ can be calculated in the same manner, except that W will now include members from the same cluster as the recipient.

5a.ii. Let $V(m_j)$ denote the set of clusters spanned by participants in a message m_j, and $V(M)$ denote the set of clusters spanned by all messages $m_j \in M$. The social network based component of completeness provided to user u_i by messages M belonging to a broad topic t_k, is denoted as $SNCompleteness_{it}$. As before, we drop the subscript t:

$$SNCompleteness_i = \sum_{j \in V(M)} |V_j|.\gamma_{ij}(W_j) \tag{8}$$

Here, W_j includes those ties that lead from V_i to V_j, among the participants in M. The summation therefore denotes the sum of the completeness contributed by individual clusters to which participants of M belong. This is intuitively equivalent to the "area" of the social network graph spanned by the messages. Fig.7.4 and 7.5 show the computation of the social network based component of completeness for a message collection across two stages.

5b. **Message length and credibility**: The completeness components based on lengths of individual contributions and the credibility of the components are calculated in the same manner as that for context (eqn. 3, 4), and are included as weights in the summation for $SNCompleteness_{_i}$ (eqn. 8) in function g described next.

$$info_j = \sum_{b \in W_j} info_{_b},\ Cred_{_j} = \sum_{b \in W_j} Cred_{_b} \tag{9}$$

5c. MO = Evidence variable for the *current amount of completeness already provided to the user*: Referring to Fig.7.3, we now express the completeness $(mo)_{_i}$ provided to user n_i by a set of messages M seen by the user:

$$(mo)_{_i} = \sum_{j \in V(M)} g(SNCompleteness_{_{ij}}, info_{_j}, Cred_{_j}) \tag{10}$$

Similar to the function f for $(na)_{_i}$, an appropriate function g for $(mo)_{_i}$ now needs to be identified, to combine the social network component with the length and credibility components. The difference is that for the calculation of completeness, these components are combined collectively for all messages $\in M$, rather than individually for each message as in the calculation for context.

5d. MA = Evidence variable for the *new amount of completeness provided by a message*: Referring to Fig.7.3 and following the same method as above, the social network component of the additional amount of completeness provided by a new message m_x to user n_i is expressed as $SNCompleteness_{_{ix}}$:

$$SNCompleteness_{_{ix}} = \sum_{j \in V(m_x) \setminus V(M)} |V_j|.\gamma_{ij}(W_j) + \sum_{j \in V(m_x) \bigcap V(M)} |V_j|.(\gamma_{ij}(W_j') - \gamma_{ij}(W_j''))$$

Here, W_j includes those users $\in V_j$ who are linked through weak ties from V_i to V_j, from among the participants in m_x. The first summation therefore denotes the completeness contributed by new clusters to which participants of m_x belong, that did not have any participation from users in the earlier messages M seen by the user. The second summation includes clusters that are common among message m_x and the earlier messages M seen by the user, but it only considers the additional amount of completeness provided by new participants in m_x who did not participate earlier in M. This is captured by considering the difference in γ calculated on W_j' and W_j'', where W_j'' includes those users $\in V_j$ who are linked through weak ties from V_i to V_j, from among participants only in $U(M)$, and W_j' includes the corresponding set of users $\in U(m_x) \bigcup U(M)$.

The overall amount of completeness provided by a new message is now be expressed by combining each Right-Hand-Side component in the summation of $SNCompleteness_{ix}$, with $info_{jx}$ and $Cred_{jx}$ using the same function g.

Freshness variables: *nf* and *mf*

6a. NF = Evidence variable for *freshness of the contextual information provided by a message*: In this thesis, we use a naive heuristic to calculate freshness. We assume that the event becomes relevant from the time instance of the first message contribution, and consider the reciprocal of the time elapsed for subsequent contributions made to the message as an measure of freshness. The contextual freshness is then calculated as the mean of the freshness of the contributions made by participants strongly linked to the recipient.

6b. MF = Evidence variable for *freshness of the completeness provided by a message*: This is calculated in the same manner as the freshness of contextual information provided by a message, except that all contributions from strong and weak ties are considered in this case.

7.1.5 Learning and inference

During the learning phase, the evidence variables **NA, NO, NF, MA, MO, MF** are calculated as described above. Knowledge of the user ratings for the usefulness variable **U** then allows us to learn the parameters for the usefulness model using standard algorithms such as EM [123].

During the inference phase, the learned usefulness model is used to estimate $P(U)$. This is calculated using standard MCMC (Markov Chain Monte Carlo) or Join-Tree belief propagation algorithms for Bayesian networks [123]. The value of $P(U = u)$ is used to decide whether the message should be recommended to the user. Algorithm-7.1 describes this process.

7.1.6 Evaluation

We now present an evaluation of the algorithm for different users in terms of the correct prediction of message usefulness ratings given by the users.

Dataset: The same **Orkut** dataset as in Chapter 6 was used to evaluate the algorithm. Users in Orkut can subscribe to communities of interest and participate in discussions in these communities. We consider a *community* equivalent to the granularity of a *broad topic* as defined earlier, a *discussion* equivalent to a *message collection* within the broad topic, and a *posting* in a discussion equivalent to a *message*. For example, a community on *Politics* (∼ broad topic) may have a discussion about *911* (∼ message collection), with many postings in the discussion (∼ messages). Orkut however does not support message ratings. We therefore recruited

Algorithm 7.1: Non-collaborative recommendation algorithm

```
Scope: Given topic = $t \in T$
1. Model learning for user $i$
Input: Message collections $M$ about events $e$ related to $t$ read by user $i$ in the past
       Participants $A(m \in M)$, Contributions $B(m \in M)$, Ratings $U_i(m \in M)$
       Topic specific social network $G_t$
Output: Learned model
forall $m \in M$ do
    Calculate $na_{im}, no_{im}, nf_{im}, ma_{im}, mo_{im}, mf_{im}$
Use EM to learn $\mathrm{P}_{it}(\mathbf{U} \mid \mathbf{NA,NO,NF,MA,MO,MF})$
2. Make recommendation decision
Input: Learned model for user $i$ and topic $t$
       New message $m$ about event $e$ related to $t$
       Participants $A(m)$, Contributions $B(m)$
Output: $\mathrm{P}_{it}(u_m \mid A(m), B(m))$          // Usefulness of message $m$ for user $i$
Calculate $na_{im}, no_{im}, nf_{im}, ma_{im}, mo_{im}, mf_{im}$
Use MCMC to infer $\mathrm{P}_{it}(u_m \mid na_{im}, no_{im}, nf_{im}, ma_{im}, mo_{im}, mf_{im})$
```

volunteers from randomly selected users in 4 communities, and asked them to rate 10 messages each in 4 message collections from the same topic. A screen-shot of the application we designed to record user ratings in shown in Appendix C. Ratings from 5 users were obtained.

Experiment: We used an open-source package, OpenBayes, to program our model. The model was simplified by discretizing the evidence variables of **NA, NO, NF, MA, MO, MF** into 3 states, the hidden variables of **NU, MU** into 2 states, and a binary classification for the usefulness variable $\mathbf{U} \in \{\text{useful, not useful}\}$. We assumed that users read the messages in order, so that we could estimate the variables in an incremental manner. For each user, the performance of the classifier was then studied with different choices of functions f and g to combine the social network based components, length, and credibility of messages. These functions can be inferred statistically in the presence of more data; in this work, we used the following functions for context f, and similar functions for completeness g:

- *Polynomial product*: $f = (SNContext)^{\{0.5,1,2\}}.(info)^{\{0.5,1,2\}}.(Cred)^{\{0.5,1,2\}}$: We studied different permutations of the exponents to examine the relative effects of the social-network component, the length, and the credibility. Within each experiment, the same permutation is used for all of **NA, NO, MA, MO**.

- *Log product*: $f = (SNContext).log_2(2 + info).log_2(2 + Cred)$: The logarithms were applied to reflect a subdued increase in the relative importance of different components.

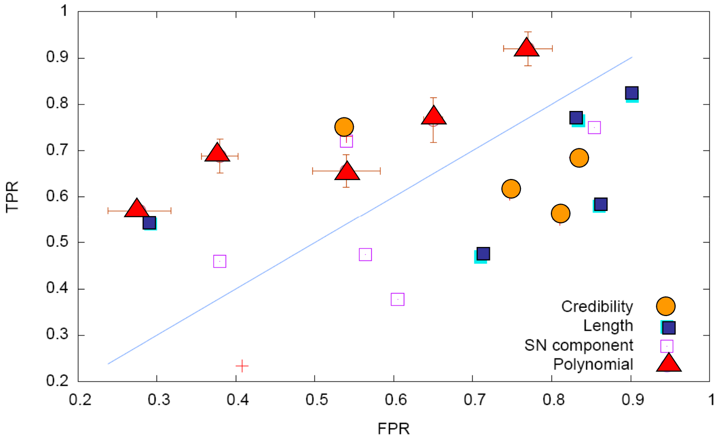


Figure 7.6: Performance across different context and completeness composition functions

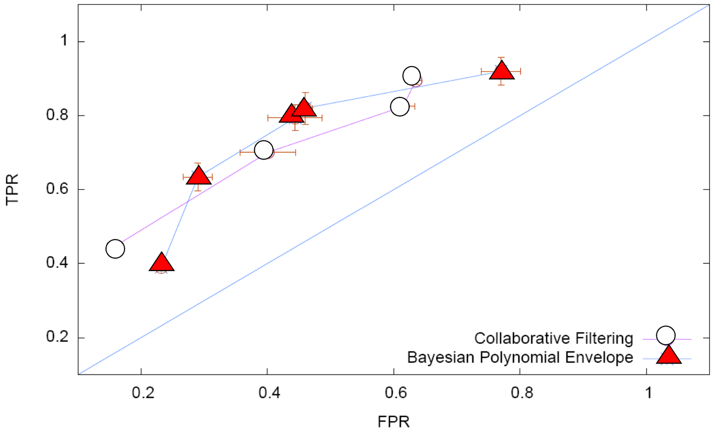


Figure 7.7: Performance comparison of the non-collaborative usefulness algorithm with CF

- *Single component*: $f = \{SNContext, info, Cred\}$: Only a single component was considered in each experiment to study its impact on performance.

For each experiment, we ran cross validation tests to study the performance of the classifier. The model was learned using 80% randomly selected ratings with the EM algorithm. The remaining 20% of ratings were inferred using the MCMC and Join-Tree implementations in OpenBayes. Both the methods gave similar results; we only show the MCMC results here. Since we are interested in the binary classification $P(U = 0, 1)$, we plot the true-positive-rate (TPR) and false-positive-rate (FPR) to test the performance [131]. Our goal is to achieve high TPR with low FPR in the classification produced by the user-model.

Results: Fig.7.6 shows the results using a polynomial function for f and g to compose context and completeness, and using each of the social network, credibility, and length components independently. Each point is the (TPR, FPR) result for a single user; since we have 5 users, there are 5 points for each function. The polynomial functions consistently dominate functions that consider only a single component, illustrating the value of considering all the components for message usefulness assessment.

Fig.7.7 shows the results for the envelope across all polynomial functions. Results using the collaborative filtering (CF) approach are also shown for comparison, but these are for only 4 users because the ratings of the 5^{th} user were not highly correlated (< 0.2 [130]) with ratings of any other user. CF is a collaborative algorithm, and will be discussed in more detail in the next section.

Although this evaluation is for only a few users, but the results are encouraging because the performance of the polynomial functions is quite close to that of the CF approach. Further, (a) our approach can produce results even for users whose preferences are not correlated with those of other users, and (b) there is much room for improvement in our results. We used very naive heuristics for length and credibility, and un-weighted clustering and integration coefficients. More sophisticated measures such as those discussed in Section 7.3 may produce better results. We next describe the collaborative algorithm, and later explain how it can be integrated with the usefulness model.

7.2 Collaborative algorithm

This algorithm works by learning a *credibility model* for each user, which indicates the degree to which a user derives credibility from the multi-dimensional bases described in Section 5.4. As discussed in Section 5.5, the models can be learned given knowledge of the social network of the users, authors of messages, and credibility ratings given by various users to these messages. Once these models have been learned for all users, the model parameters are used to infer sets of users having similar model characteristics. Then credibility ratings given to new messages by similar users are used to predict the credibility of the messages for a given recipient user, much like collaborative filtering [120]. However, unlike CF which only reproduces average behavior similarity, our approach uses four sets of similarity metrics between users, and is hence able to produce closer fitted recommendations. We next explain the rationale behind the multi-dimensional credibility model, and then describe the model and our evaluation results in detail.

Note that although the algorithm is designed to work on the level of individual contributions to a message, but the dataset used for validation consists of only a single story per message without any comments; therefore, for ease of exposition, we operate at the *message* level, rather than with contributions.

Building upon the observations made in Chapter 5, we state the following expressibility goals that our credibility model should be able to capture:

- *E-1*: Different users may judge the credibility of a messages differently according to their own context.

- *E-2*: Different users may associate different degrees of credibility to the public opinion or to the beliefs of other groups of users or to their own beliefs.

- *E-3*: The credibility of an user is topic specific; an expert in some area may not be an expert in another.

- *E-4*: A highly credible user can occasionally make mistakes and give inaccurate information. Analogously, useful messages could be written by a user new to the recommender system.

We are able to meet these expressibility goals by modeling criteria discussed next, such as the influence of public opinion on users, the influence of close friends, and the extent to which different users may be willing to diverge from their own preconceived beliefs. These criteria are used to build and learn a Bayesian network for each user, to predict which messages the user may find credible.

7.2.1 Credibility judgement criteria

Multi-dimensional construct: Various researchers have proposed to model credibility as a multi-dimensional construct [102,103]. [102] reason about credibility criteria used by people to judge the credibility of computerized devices and software, and identify the following different types of credibility:

- *Experienced*: This is based on *first-hand experience* of a user, and reflects her personal belief about a device or software.

- *Presumed*: This reflects personal biases of a user that give rise to *general assumptions* about certain categories of computing products; for example, presumptions based upon the company which developed the product, the cost of the product, the importance of the function performed by the product, etc.

- *Reputed*: This is based on reports about different products from *third-party* sources.

A model with similar distinctions is developed in [103] to evaluate the trustworthiness of agents in an e-commerce setting. Here, the authors distinguish *witness reputation* (i.e. general public opinion) from *direct reputation* (i.e. opinion from a user's own experience) and include as well *system reputation* (i.e. the reputation from the role of an agent, as buyer, seller or broker). We next consider relevant studies from sociology and political science for additional valuable insights.

Social networks: One among many studies based on the *strength-of-weak-ties* hypothesis, [98] traces the changes in political opinion of people before and after the 1996 presidential elections in USA, observed with respect to the social networks of people. It is shown that weak ties (identified as geographically dispersed ties

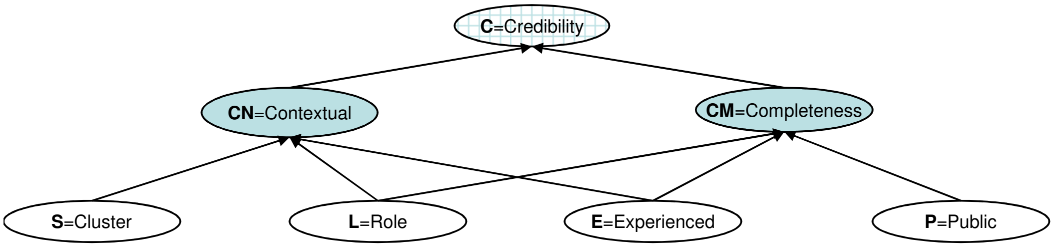


Figure 7.8: Credibility model

of acquaintances) are primarily responsible for the diffusion of divergent political opinion into localized clusters of people having strong ties between themselves. This reflects that local community clusters of people are often homogeneous in opinion, and these opinions may be different from those of people belonging to other clusters. Furthermore, people may to different degrees respect opinions different from those of their immediate local community members. This reflects that the personal characteristics of people also influence the extent to which they would be comfortable in deviating from the beliefs of their immediate local cluster. These observations provide two insights:

- Reputed credibility has at least two sub-types: *cluster credibility* based on the opinions of people in the same cluster or local community, and *public credibility* based on the general opinions of everybody.
- Users have different personal characteristics to weigh the importance of these two types of credibilities.

The first insight suggests refining *reputed* credibility to also consider reports from those in the same cluster. The second insight is reinforced by studies in information science [91], which argue that users have different preferences for different types of credibilities discussed so far. Inspired by these studies, we develop and operationalize a multi-dimensional subjective credibility model for participatory media as described next.

7.2.2 Credibility model

Knowledge assumptions: Suppose that we wish to predict whether a message m_k about a topic t and written by user n_j, will be considered credible by user n_i. As before, we assume that we have the following prior knowledge:

- The topic specific social network graphs formed by all users and the links between pairs of users.

- Classification of ties between users as strong or weak. We use V_{it} to denote the local cluster of users strongly tied to user n_i with respect to topic t.

- Authors of all messages about all topics.

- Credibility ratings assigned by various users to messages [2].

These assumptions are reasonable in contexts such as social networking and knowledge sharing websites such as Digg and LiveJournal, which allow users to construct social networks, post messages, and assign ratings to messages. We will show that we can use this knowledge to quantify different types of credibilities for each message with respect to each user. Then, based on ratings given by a particular user to messages seen in the past, we use a Bayesian model to learn preferences of the user towards these different kinds of credibilities of messages. Finally, we use this learned model to predict if the new message m_k will be considered credible by user n_i.

Bayesian network: The different types of credibilities that we model are as follows:

- e_{ikt}= *experienced credibility*: Identical to the concept of experienced credibility discussed earlier, this is based only on ratings given by user n_i in the past, and denotes the credibility that n_i associates with the message m_k written by n_j.

- l_{ikt}= *role based credibility*: Similar to presumed credibility discussed earlier, this denotes the credibility that n_i associates with the message m_k written by users having the same role as that of n_j; for example, based on whether the messages' authors are students, or professors, or journalists, etc.

- s_{ikt} = *cluster credibility*: A sub-type of reputed credibility discussed earlier, this is based on the ratings given by other users in cluster V_{it}, that is, the cluster of user u_i. It denotes the credibility associated by the cluster or local community of n_i with the message m_k written by n_j.

- p_{kt} = *public credibility*: Another sub-type of reputed credibility, this is based on ratings by all the users, and reflects the general public opinion about the credibility for the message m_k written by n_j.

Each of these credibilities can be expressed as a real number in the unit interval, and we propose a Bayesian network to combine them into a single credibility score. The model is shown in Fig.7.8. Our aim is to learn the distribution for $\mathrm{P}_{it}(\mathbf{C} \mid \mathbf{E},\mathbf{L},\mathbf{S},\mathbf{P})$ for each user and topic based on ratings given by various users to existing messages; here, $\{\mathbf{E},\mathbf{L},\mathbf{S},\mathbf{P}\}$ are evidence variables for the four types of credibilities

[2] We assume that we are beyond the cold-start stage so that all messages have received some ratings, and all users have provided at least some ratings.

for a message, and **C** is a variable denoting the credibility that n_i associates with the message. Thus, for each topic t, a set of messages M about t will be used during the training phase with samples of (c_{ik}, e_{ik}, l_{ik}, s_{ik}, p_k) for different messages $m_k \in M$ to learn the topic specific credibility models for n_i.

Fig.7.8 also shows two hidden variables as shaded ovals for contextual and completeness credibility of messages. The hidden variables help make the model more tractable to learn, and also directly model context and completeness which makes the credibility model compatible with the usefulness model, described in Section 7.3. For each message, the model first estimates the credibilities of the contextual and complete information carried by the message, and then uses these two credibilities to generate the final estimate. We reason that cluster credibility will only influence contextual credibility, while public credibility will only influence completeness credibility. This is because general public opinion is by definition averaged over different contexts, and hence it will only add noise to any context specific credibility. Similarly, cluster credibility will double count the opinion of a specific cluster when judging the degree of completeness or diversity in a message. Other types of credibilities, experienced and role based, will influence both contextual and completeness credibility since they are based on the personal beliefs of the user.

Meeting the expressibility goals: Our modeling method satisfies three out of the four expressibility goals. (*E-1*) The model takes into account personal and contextual opinions of people that may influence their credibility judgements. (*E-2*) The model is learned for each user, and allows varying degrees of openness of users to respect opinions of other users. (*E-3*) Different model instances are learned for different topics, making credibility judgements topic specific. (*E-4*) We will show later in Section 7.2.4 that the fourth goal of allowing mistakes by credible users and useful messages by less-credible users can also be modeled in this framework.

7.2.3 Credibility computation

In this section, we describe how the different types of credibilities can be computed based on social network information, ratings given by users to messages, and authorship information. The notion of credibility of messages is extended to credibility of users as well. We first list the rules that are the basis for our formulation to quantify the various types of credibilities, and then give the actual computation process.

Rules to calculate credibility

We use the information captured in the following relationships, constrained primarily by our knowledge assumptions, to meet the expressibility goals outlined earlier:

- *R-1*: A message is credible if it is rated highly by credible users.

- *R-2*: A user is credible if messages written by her are rated highly by other credible users.
- *R-3*: A user is also credible if ratings given by her are consistent with the ratings given by credible users.
- *R-4*: A user is also credible if she is linked to by other credible users in the social network.

There is clearly a recursive relationship between these rules. We solve the recursion using fixed-point Eigenvector computations, as described next.

Calculation of evidence variables

As stated in the knowledge assumptions earlier, we start with the following information that will be a part of our training set for topic t.

- $\mathbf{A}_t$**[k,n]**: A matrix for k messages and n users, where $a_{ij} \in \{0, 1\}$ indicates whether message m_i was written by n_j
- $\mathbf{C}_t$**[k,n]**: A ratings matrix for k messages and n users, where $c_{ij} \in \{0, 1\}$ [3] indicates the rating given to message m_i by user n_j
- $\mathbf{N}_t$**[n,n]**: A social network matrix where $n_{ij} \in \{0, 1\}$ indicates the presence or absence of a link from user n_i to user n_j. This is the same as the topic specific social network graph G_t introduced in Chapter 6. We also assume that the clustering algorithm can identify clusters of strong ties among users, connected to other clusters through weak ties.

Our goal is to find a method to compute the evidence variables for the Bayesian model using the rules given above. The evidence variables can be expressed as the matrices $\mathbf{E}_t$**[n,k]**, $\mathbf{L}_t$**[n,k]**, $\mathbf{S}_t$**[n,k]**, and $\mathbf{P}_t$**[k]**, containing the credibility values for messages. Here, p_k is the public credibility for message m_k authored by user n_j. e_{ij} and l_{ij} are the experienced and role based credibilities respectively for message m_k according to the self-beliefs of user n_i. Similarly, s_{ij} is the cluster credibility for message m_k according to the beliefs of the users in n_i's cluster V_i. Once these evidence variables are computed for existing messages, they are used to learn the Bayesian model for each user. Subsequently, for a new message, the learned model for a user is used to predict the credibility of the new message for the user.

We henceforth assume that we are operating within some topic t, and drop the subscript. We begin with computation of the evidence variable matrix for public credibility $\mathbf{P}$, and will explain later how other credibilities can be computed in a similar fashion.

[3] We assume in this thesis that the ratings are binary. However, our method can be easily generalized to real-valued ratings as well. In the future, we also plan to extend the method to accept explicit negative ratings using distrust propagation [143].

1. Let $\mathbf{P'[n]}$ be a matrix containing the public credibilities of users, and consider the credibility of a message as the mean of the ratings for the message, weighted by the credibility of the raters (*r-1*):

   $$p_k = \sum_i c_{ki}.p'_i/|c_{ki} > 0|$$

   Here, the denominator counts the number of occurrences of ratings greater than 0. This is the same as writing $\mathbf{P}$=$\mathbf{C}_r.\mathbf{P'}$, where $\mathbf{C}_r$ is the row-stochastic form of $\mathbf{C}$, ie. the sum of elements of each row = 1.

2. The credibility of users is calculated as follows:

2a. Consider the credibility of a user as the mean of the credibilities of the messages written by her (*R-2*):

$$p'_i = \sum_k p_k/|p_k|$$

This is the same as writing $\mathbf{P'}$=$\mathbf{A}_c^T.\mathbf{P}$, where $\mathbf{A}_c$ is the column-stochastic form of $\mathbf{A}$; and $\mathbf{A}_c^T$ is the transpose of $\mathbf{A}_c$.

2b. The above formulation indicates a fixed point computation:

$$\mathbf{P'}=\mathbf{A}_c^T.\mathbf{C}_r.\mathbf{P'} \qquad (1)$$

Thus, $\mathbf{P'}$ can be computed as the dominant Eigenvector of $\mathbf{A}_c^T.\mathbf{C}_r$. This formulation models the first two rules, but not yet the ratings-based credibility (*R-3*) and social network structure of the users (*R-4*). This is done as explained next.

2c. Perform a fixed-point computation to infer the credibilities $\mathbf{G[n]}$ acquired by users from the social network (*R-4*):

$$\mathbf{G}=(\beta.\mathbf{N}_r^T + \mathbf{(1\text{-}}\beta).\mathbf{Z}_c.\mathbf{1}^T).\mathbf{G} \qquad (2)$$

Here, $\beta \in (0,1)$ denotes a weighting factor to combine the social network matrix $\mathbf{N}$ with the matrix $\mathbf{Z}$ that carries information about ratings given to messages by users. We generate $\mathbf{Z}$ by computing z_i as the mean similarity in credibility ratings of user n_i with all other users, as in the CF algorithm. The ratings similarity between a pair of users is computed as the **Jacquard's coefficient** of common ratings between the users. Thus, z_i will be high for users who give credible ratings, that is, their ratings agree with the ratings of other users (*R-3*). In this way, combining the social-network matrix with ratings-based credibility helps to model the two remaining rules as well. Note that $\mathbf{Z}_c\mathbf{[n]}$ is a column stochastic matrix and $\mathbf{1[n]}$ is a unit column matrix; augmenting $\mathbf{N}$ with $\mathbf{Z}_c.\mathbf{1}^T$ provides an additional benefit of converting $\mathbf{N}$ into an irreducible matrix so that its Eigenvector can be computed [4].

[4] This step is similar to the Pagerank or HITS computations for the importance of Internet

2d. The ratings and social network based scores are then combined together as:

$$\mathbf{P'}\text{=}(\alpha.\mathbf{A}_c^T.\mathbf{C}_r + \mathbf{(1\text{-}}\alpha\mathbf{)}.\mathbf{G}_c.\mathbf{1}^T).\mathbf{P'} \tag{3}$$

Here again $\mathbf{1}$ is a unit column matrix, and $\alpha \in (0, 1)$ is a weighting factor. The matrix $\mathbf{P'}$ can now be computed as the dominant Eigenvector using the power method [122].

3. Once $\mathbf{P'}$ is obtained, $\mathbf{P}$ is calculated in a straightforward manner as $\mathbf{P}\text{=}\mathbf{C}_r.\mathbf{P'}$.

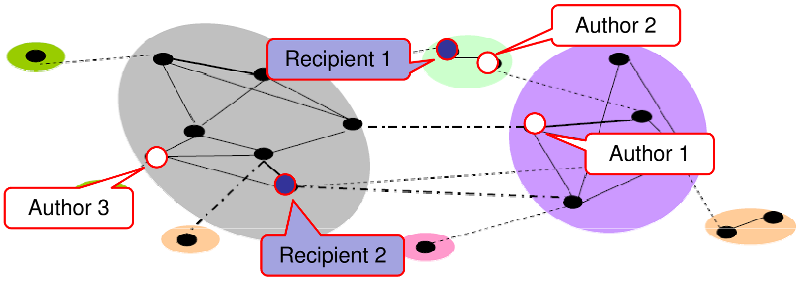
Note that the above method is only one way of combining the different pieces of information we have. Our objective in presenting this method is to show that information about social networks, ratings, and authorship can be mathematically integrated with well-known ranking techniques, and to then examine the performance of this method compared to competing approaches.

The above process is to compute the public credibilities **P[k]** of messages. The processes to compute cluster **S[n,k]**, experienced **E[n,k]**, and role based **L[n,k]** credibilities are identical, except that different cluster credibilities are calculated with respect to each cluster in the social network, and different experienced and role based credibilities are calculated with respect to each user. This is why cluster and experienced credibility matrices are 2-dimensional, while the public credibility is only 1-dimensional. For example, considering a message m_3 and a recipient user n_1, **P[3]** is the public credibility of message m_3; **E[1,3]** is the experienced credibility of message m_3 according to the self-belief of recipient n_1; **L[1,3]** is the role based credibility of message m_3 also according to the self-belief of recipient n_1; and **S[1,3]** is the cluster credibility of message m_3 according to the beliefs of users in cluster V_1 of recipient n_1. The processing steps for computing these quantities are outlined in Algorithm-7.2; a description is below.

- The cluster credibilities **S[n,k]** are computed in the same manner as the public credibilities, but after modifying the ratings matrix **C** to contain only the ratings of members of the same cluster. Thus, the above process is repeated for each cluster, modifying **C** in every case. For each users n_i belonging to cluster V_i, s_{ik} is then equal to the cluster credibility value for message m_k with respect to n_i.

  The matrix **Z** in the computation on the social network matrix is also modified. When computing the cluster credibilities for cluster V_i, element z_j of **Z** is calculated as the mean similarity of user n_j with users in cluster V_i. Thus, z_j will be high for users who are regarded credible by members of cluster V_i because their ratings agree with the ratings of the cluster members.

web pages [132,133]. The matrix **N** can be considered as the link matrix of web-pages, and the matrix **Z** as the pagerank personalization matrix. The output matrix **G** then essentially ranks the web-pages in order of their importance, after taking personalization into account.



	Context r_2	Completeness r_2
Context r_1	Contextual similarity	Msgs by author 2
Completeness r_1	Msgs by author 3	Msgs by author 1

Figure 7.9: Context and completeness similarity between users

- The experienced credibilities **E**[**n**,**k**] are computed in the same manner as well, but this time for each user by modifying the ratings matrix **C** to contain only the ratings given by the user. The matrix **Z** is also modified each time by considering z_j as the similarity between users n_i and n_j, when calculating the experienced credibilities for n_i.

- Role based credibility is computed as the mean experienced credibilities of users having the same role. However, we do not use role based credibility in our evaluation because sufficient user profile information was not available in the experimental dataset used by us. Henceforth, we ignore **L**[**n**,**k**] in our computations.

7.2.4 Learning and inference

Once the various types of credibilities for messages are calculated with respect to different users, this training data is used to learn the Bayesian model for each user and topic of interest to the user using the Expectation-Maximization (EM) algorithm [123]. At this point, the learned model can be used in two ways. First, as with the usefulness model, the credibility model can be used to infer $\mathrm{P}_{it}(\mathbf{C} \mid \mathbf{E},\mathbf{S},\mathbf{P})$ for new messages to predict whether user n_i will find new messages about topic t to be credible. We experimented with this method in [193, 194], and details are given in Appendix D, but this method is structurally similar to the non-collaborative method described in the Section 7.1 because it uses the learned model to directly make a prediction about message credibility for user n_i. Here, we describe an alternative method, which follows the technique of collaborative recommendation algorithms.

Algorithm 7.2: Credibility model: Training set preparation

```
Scope: Given topic = $t \in T$
Input: $\mathbf{A[k,n]}$, $\mathbf{C[k,n]}$, $\mathbf{N[n,n]}$;        // Authorship, ratings, and social network matrices
Output: $\mathbf{P[k]}$, $\mathbf{E[n,k]}$, $\mathbf{S[n,k]}$, $\mathbf{P'[k]}$, $\mathbf{E'[n,n]}$, $\mathbf{S'[n,n]}$;        // Message and user cred

1. Compute similarity matrix $\mathbf{Y[n,n]}$
forall $i \in 1..n,\ j \in 1..n,\ i \neq j$ do
    forall $m \in 1..k$ do
        if $C[m,i] = C[m,j]$ then
            $\mathbf{Y[i,j]} \leftarrow \mathbf{Y[i,j]}+\frac{1}{k}$;        // Similarity between pairs of users

2. Compute public credibilities $\mathbf{P[k]}$, $\mathbf{P'[n]}$
$\mathbf{Z[n]} \leftarrow 0$
forall $i \in 1..n$ do
    forall $j \in 1..n$ do
        $\mathbf{Z[i]} \leftarrow \mathbf{Z[i]} + \mathbf{Y[j,i]}$;        // Agreement of user's ratings with other users
Solve for $\mathbf{G[n]}$: $\mathbf{G}=(\beta.\mathbf{N}_r^T+(\mathbf{1}\text{-}\beta).\mathbf{Z}_c.\mathbf{1}^T).\mathbf{G}$;        // Use power method
Solve for $\mathbf{P'[n]}$: $\mathbf{P'}=(\alpha.\mathbf{A}_c^T.\mathbf{C}_r + (\mathbf{1}\text{-}\alpha).\mathbf{G}_c.\mathbf{1}^T).\mathbf{P'}$;        // Use power method
$\mathbf{P} \leftarrow \mathbf{C}_r.\mathbf{P'}$

3. Compute cluster credibilities $\mathbf{S[n,k]}$, $\mathbf{S'[n,n]}$
forall $Cluster\ V_c \in clusters\ in\ social\ network$ do
    $\mathbf{Z[n]} \leftarrow 0$
    $\underline{\mathbf{G}}\mathbf{[n]} \leftarrow 0$, $\underline{\mathbf{P}}\mathbf{[n]} \leftarrow 0$, $\underline{\mathbf{P'}}\mathbf{[n]} \leftarrow 0$, $\underline{\mathbf{C}}\mathbf{[k,n]} \leftarrow 0$;        // Scratch variables
    forall $j \in users\ in\ V_c$ do
        forall $i \in 1..n$ do
            $\mathbf{Z[i]} \leftarrow \mathbf{Z[i]} + \mathbf{Y[j,i]}$;    // Agreement of user's ratings with cluster members
        forall $m \in 1..k$ do
            $\underline{\mathbf{C}}\mathbf{[m,j]} \leftarrow \mathbf{C[m,j]}$;        // Use only ratings of cluster
            members
    Solve for $\underline{\mathbf{G}}\mathbf{[n]}$: $\underline{\mathbf{G}}=(\beta.\mathbf{N}_r^T+(\mathbf{1}\text{-}\beta).\mathbf{Z}_c.\mathbf{1}^T).\underline{\mathbf{G}}$;        // Use power method
    Solve for $\underline{\mathbf{P}}'\mathbf{[n]}$: $\underline{\mathbf{P}}'=(\alpha.\mathbf{A}_c^T.\underline{\mathbf{C}}_r+(\mathbf{1}\text{-}\alpha).\underline{\mathbf{G}}_c.\mathbf{1}^T).\underline{\mathbf{P}}'$;        // Use power method
    $\underline{\mathbf{P}}=\underline{\mathbf{C}}_r.\underline{\mathbf{P}}'$
    forall $j \in users\ in\ V_c$ do
        forall $m \in 1..k,\ u \in 1..n$ do
            $\mathbf{S'[j,u]} \leftarrow \underline{\mathbf{P'}}\mathbf{[u]}$; $\mathbf{S[j,m]} \leftarrow \underline{\mathbf{P}}\mathbf{[m]}$

4. Compute experienced credibilities $\mathbf{E[n,k]}$, $\mathbf{E'[n,n]}$
forall $User\ i \in 1..n$ do
    $\mathbf{Z[n]} \leftarrow 0$
    $\underline{\mathbf{G}}\mathbf{[n]} \leftarrow 0$, $\underline{\mathbf{P}}\mathbf{[n]} \leftarrow 0$, $\underline{\mathbf{P'}}\mathbf{[n]} \leftarrow 0$, $\underline{\mathbf{C}}\mathbf{[k,n]} \leftarrow 0$
    forall $j \in 1..n$ do
        $\mathbf{Z[j]} \leftarrow \mathbf{Y[j,i]}$;        // Agreement of user's ratings with recipient
    forall $m \in 1..k$ do
        $\underline{\mathbf{C}}\mathbf{[m,i]} \leftarrow \mathbf{C[m,i]}$;        // Use only ratings of recipient
    Solve for $\underline{\mathbf{G}}\mathbf{[n]}$: $\underline{\mathbf{G}}=(\beta.\mathbf{N}_r^T+(\mathbf{1}\text{-}\beta).\mathbf{Z}_c.\mathbf{1}^T).\underline{\mathbf{G}}$;        // Use power method
    Solve for $\underline{\mathbf{P}}'\mathbf{[n]}$: $\underline{\mathbf{P}}'=(\alpha.\mathbf{A}_c^T.\underline{\mathbf{C}}_r+(\mathbf{1}\text{-}\alpha).\underline{\mathbf{G}}_c.\mathbf{1}^T).\underline{\mathbf{P}}'$;        // Use power method
    $\underline{\mathbf{P}} \leftarrow \underline{\mathbf{C}}_r.\underline{\mathbf{P}}'$
    forall $m \in 1..k,\ u \in 1..n$ do
        $\mathbf{E'[i,u]} \leftarrow \underline{\mathbf{P'}}\mathbf{[u]}$; $\mathbf{E[i,m]} \leftarrow \underline{\mathbf{P}}\mathbf{[m]}$
```

Once the model parameters are learned for each user, MCMC or other belief propagation methods are used to infer $\mathrm{P}_{it}(\mathbf{CN} \mid \mathbf{C,E,S,P})$ and $\mathrm{P}_{it}(\mathbf{CM} \mid \mathbf{C,E,S,P})$ for the hidden variables of **CN** and **CM** as the preferences of user n_i and topic t for contextual and complete information carried by messages. At this point, we use an algorithm very similar to collaborative filtering (CF) [120].

The basic CF algorithm computes a similarity measure between a pair of users based on the similarity of message ratings given by each pair. Then, a decision is made whether a new message should be recommended to a recipient user by calculating the mean of the ratings given to the message by other users similar to the recipient user. If the mean is greater than a threshold, the message is recommended; else it is rejected. The drawback of the CF method is that it only learns the average user behavior. But since we have two parameters **CN** and **CM** to describe a user's behavior towards each message, we can compute more granular similarities between each pair of users, as follows.

Correlating pairs of users on instances of **CN** over the set of messages in the training set gives information about the contextual similarity between users. In fact, we calculate four sets of similarities between each pair of users based on whether messages are believed to be highly contextual by both users (**CN**, **CN**), or highly complete by both users (**CM**, **CM**), or contextual by the first user and complete by the second user (**CN**, **CM**), or vice versa (**CM**, **CN**). This is shown schematically in Fig.7.9. The figure shows two recipient users in adjacent clusters, and authors in the same or different clusters. Carrying forward the insight from Chapter 5 that users in the same social network cluster tend to share the same context, messages by different authors can be categorized in terms of the 2 X 2 table for context and completeness similarity between the two recipients r_1 and r_2. Thus, four sets of similarities between any pair of users can be computed. This provides two advantages. First, since this method breaks down the user behavior into four components based upon the context and completeness of messages to users, it can produce a closer fitting to the user behavior than CF which is only based on the average user behavior. Second, the method can operate across non-adjacent clusters as well, and gets around the limitation of the usefulness model used in the non-collaborative algorithm as being restricted to only adjacent clusters. The processing steps for producing recommendations are outlined in Algorithm-7.3; a description is below:

1. For each user, run the EM algorithm on the training set to learn the model.

2. Use the learned model to infer the probabilities of the hidden variables of context and completeness for each story in the training set: $\mathrm{P}_i(\mathbf{CN}|\mathbf{E,S,P,C})$ and $\mathrm{P}_i(\mathbf{CM}|\mathbf{E,S,P,C})$ shown in Fig.7.8. That is, for each story m_j, infer $\mathrm{P}(cn_{ji} \,|e_{ji}, s_{ji}, p_{ji}, c_{ji})$ and $\mathrm{P}(cm_{ji} \,|e_{ji}, s_{ji}, p_{ji}, c_{ji})$.

3. Discretize the probabilities for **CN** and **CM** into a binary class. This gives samples of ($c_{ji} \in \{0,1\}$, $cn_{ji} \in \{0,1\}$, $cm_{ji} \in \{0,1\}$), that is, which stories

appear contextual or complete to a user, and the rating given by the user to these stories. The discretization is done by comparing the distribution of **CN** and **CM** for each user with respect to the ratings **C** given by the user, to find a *context threshold* and *completeness threshold*, so that instances of cn_{ji} below the *context threshold* are discretized to 0, instances of cn_{ji} above the context threshold are discretized to 1, and similarly for the discretization of cm_{ji}. This is described in more detail in the evaluation section.

4. For every pair of users, compare the samples to produce four similarity coefficients on how similar the users are in their contextual opinion, completeness opinion, and cross opinions between messages that appear contextual to one user and complete to the other, or vice versa.

5. Finally, to evaluate the decision of whether to recommend a new message to a user, compute the weighted mean of the message ratings over all the four coefficients of similarity, rather than over a single coefficient as in the basic CF algorithm. We use two thresholds here: A *similarity threshold* to determine whether a pair of users are similar to each other on some similarity coefficient, and a *recommendation threshold* to determine whether the weighted mean is sufficiently high to recommend the message to the user.

Note that using ratings to evaluate the recommendation decision also meets the fourth expressibility goals of allowing mistakes by credible users (*E-4*) listed earlier. It allows new users to popularize useful messages written by them because their own credibility does not play a role in the computations. It also allows credible users to make mistakes because the credibility of the author is not taken into account. We next present results from evaluation of the collaborative algorithm. Later, in Section 7.3, we will outline how the non-collaborative and collaborative algorithms can be combined together to incorporate all of context, completeness, and credibility estimation in a hybrid recommendation algorithm.

7.2.5 Evaluation

Dataset: We evaluate the algorithm over a dataset of ratings by real users obtained from a popular knowledge sharing website, **digg.com** [134]. The website allows users to link to other users in the social network, and to submit links to a common pool about interesting news articles or blogs. These links are called *stories*; any Digg user can vote for these stories, known as *digging* the stories. Stories that are *dugg* by a large number of users are promoted to the front-page of Digg. Note that these *diggs* may not reflect credibility ratings, but rather usefulness ratings given to stories; we provide a justification below about why we consider them as credibility ratings. The dataset provides us with all the information we need:

- Social network of users: We use this information to construct the social network link matrix between users **N[n,n]**. The social network is clustered using

Algorithm 7.3: Collaborative recommendations

```
Scope: Given topic = $t \in T$ Input: Learned models for all users;
Output: Preprocessing required for recommendation algorithm;

1. Infer and discretize CN, CM;
forall $i \in 1..n$ do
    forall $m \in 1..k$ do
        Infer $cn_{im} \leftarrow$ P(CN = $cn$| E[i,m],C[i,m],P[m]), $cm_{im} \leftarrow$ P(CM = $cm$|
        E[i,m],C[i,m],P[m])
        // Discretize on context threshold for $i$
        CN[i,m] $\leftarrow$ discretize($cn_{im}$, $i$, context)
        // Discretize on completeness threshold for $i$
        CM[i,m] $\leftarrow$ discretize($cm_{im}$, $i$, completeness)

2. Compute similarity matrices Y[n,n,4];
forall $i \in 1..n,\ j \in 1..n,\ i \neq j$ do
    forall $m \in 1..k$ do
        if CN[i,m] = CN[j,m] then
            Y[i,j,0] $\leftarrow$ Y[i,j,0]+$\frac{1}{k}$;        // Context-Context similarity
        if CN[i,m] = CM[j,m] then
            Y[i,j,1] $\leftarrow$ Y[i,j,1]+$\frac{1}{k}$;        // Context-Completeness similarity
        if CM[i,m] = CN[j,m] then
            Y[i,j,2] $\leftarrow$ Y[i,j,2]+$\frac{1}{k}$;        // Completeness-Context similarity
        if CM[i,m] = CM[j,m] then
            Y[i,j,3] $\leftarrow$ Y[i,j,3]+$\frac{1}{k}$;        // Completeness-Completeness similarity

3. Make recommendation decision;
Input: User $i$, Message $m$;
       Ratings C[n,m] given by other users to $m$
Output: User $i$ will find $m$ to be credible?

$sum \leftarrow 0$, $count \leftarrow 0$                          // Scratch variables
forall $j \in 1..n,\ j \neq i$ do
    forall $k \in 0..3$ do
        if Y[i,j,k] > Similarity threshold then
            $sum \leftarrow sum +$ C[j,m] . Y[i,j,k]
            $count \leftarrow count +$ Y[i,j,k]
$avg \leftarrow \frac{sum}{count}$                               // Weighted mean of ratings
if $avg$ > Recommendation threshold then
    Recommend $m$ to $i$
else
    Do not recommend $m$ to $i$
```

MCL, a graph clustering algorithm [110], to produce classifications of ties as strong or weak [191]. The cluster of users strongly connected to user u_i is referred to as V_i.

- Stories submitted by various users: We use this information to construct the authorship matrix **A[k,n]**. Since all the stories in the dataset were related to technology, we consider all the stories as belonging to a single topic.
- Stories dugg by various users: We use this information to construct the ratings matrix **R[k,n]**. We consider a vote of 1 as an evidence for credibility of the story.

Although the dataset is quite large with over 200 stories, we are able to use only 85 stories which have a sufficiently large number of ratings by a common set of users. This is because we require the same users to rate many stories so that we have enough data to construct training and test datasets for these users. Eventually, we assemble a dataset of 85 stories with ratings by 27 users. We do not include users who rate more than 65 stories as all 1 or 0, because a good predictor for such users would trivially be to always return 1 or 0, and besides, such user behavior may amount to attacks on the system which we consider as future work. A few assumptions we make about the validity of the dataset for our experiments are as follows:

- The original submission of a story to Digg may not necessarily be made by the author of the story. However, we regard the submitting user as the message author. This is justified because it distinguishes this user from other users who only provide further ratings to the messages.
- The ratings provided on the Digg website may not reflect credibility ratings, but rather usefulness ratings given to messages by users. We however consider them to be equivalent to credibility because of the smaller dataset size we use. We argue that since the users in the dataset vote for at least 20 stories out of 85 (25% of the total number of stories), they are likely to be interested in the topic and all the stories; therefore, the only reason for their not voting for a story would be its credibility. Later in Section 7.3, we clarify this when we combine the credibility and usefulness models into a single comprehensive model. We will try to assemble a more extensive dataset in the future exclusively focused on credibility.

Experiment: We use an open-source package, OpenBayes, to program the Bayesian network. We simplify the model by discretizing the evidence variables **E,S,P** into 3 states, and a binary classification for the hidden variables **CN, CM**, and the credibility variable **C**. The discretization of the evidence variables into 3 states is performed by observing the Cumulative Distribution Frequency (CDF) and Complementary Cumulative Distribution Frequency (CCDF) of each variable with respect to the credibility rating of users. The lower cutoff is chosen such that the product of the CDF for rating=0 and CCDF for rating=1 is maximum, indicating the point at which the evidence variable has a high probability of being 0 and a low probability of being 1. Similarly, the upper cutoff is chosen such that the CCDF for rating=0 and CDF for rating=1 is maximum, indicating the point at which the evidence variable has a low probability of being 0 and a high probability of being 1. This gives a high discrimination ability to the classifier because the cutoffs are selected to maximize the pair-wise correlation of each evidence variable with the credibility rating given by the user. Discretization of **CN** and **CM** is performed in a similar manner by comparing the CDFs for rating=0 and CCDFs for rating=1 for both **CN** and **CM**.

We evaluate the performance of the model for each user by dividing the 85 stories into a training set of 67 stories and a test set of 17 stories (80% and 20% of the dataset respectively). We then repeat the process 20 times with different random selections of stories to get confidence bounds for the cross validation. For each evaluation, we use two kinds of performance metrics [131]:

- *Matthew's correlation coefficient*: This is computed as follows:

  $$\text{MCC} = \frac{(t_p.t_n - f_p.f_n)}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}}$$

  Here, f_p = false positives, t_p = true positives, f_n = false negatives, t_n = true negatives. The MCC is a convenient measure because it gives a single metric for the quality of binary classifications.

- *TPR-FPR plot*: This plots on an XY-scale the true positive rate (TPR) with the false positive rate (FPR) of a binary classification. The point of maximum accuracy means a TPR=1.0 and FPR=0.0, while a random baseline means TPR=FPR. Therefore, points above the random baseline are considered to be good.

Through some preliminary experiments we were able to find good values of α (eqn. 3) and β (eqn. 2) that gave the best performance on the MCC metric. We use $\alpha = 0.5$ and $\beta = 0.85$, conveying our message that all of authorship, ratings, and social networks provide valuable credibility information.

Results: We used the following threshold values: *similarity threshold* = 0.6, *recommendation threshold* = 0.5. Fig.7.10 shows the performance of the basic CF scheme and our enhanced version. The basic CF scheme performs worse than random for many users, but when enhanced with breaking up the average user behavior into contextual and completeness components, the performance improves considerably. The mean MCC for the basic scheme is 0.017 (σ = 0.086), and for the enhanced scheme is 0.278 (σ = 0.077), a sixteen-fold improvement. We consider this to be a huge improvement over the existing methodologies for trust, reputation, and recommendation algorithms, especially to build applications related to participatory

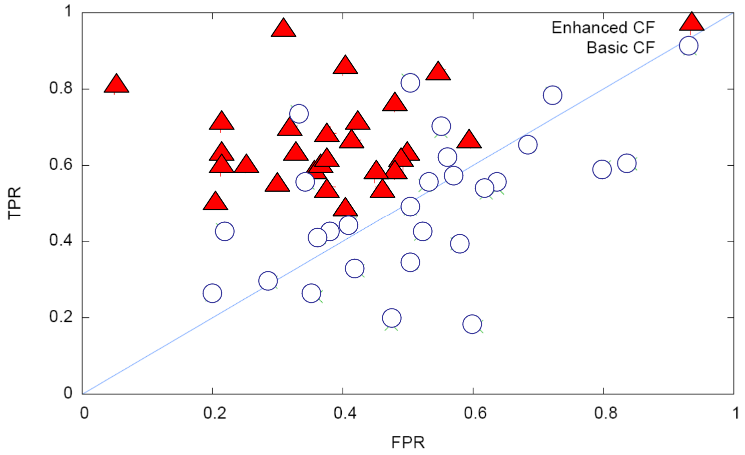


Figure 7.10: Enhancement of collaborative filtering

media. Our results reinforce the value of using sociological insights in recommender system design.

We notice that the classifier performs very well for some users, but close to random for some other users. We therefore investigate various characteristics that may prove useful to determine for which users our method may work well and when it may not.

- We compute the variance of cluster and experienced credibility scores for different users. We then compare the variances by users for whom our method works well ($\frac{TPR}{FPR} > 1.5$), with the variances by the remaining users. We find that for both cluster and experienced credibilities, the variances by users for whom our method works well are more than twice the variances by other users.

  This shows the more the discrimination produced in the cluster and experienced credibility scores by a user, the better the performance for the user, because greater discrimination ability implies higher entropy in the information theoretic sense.

- We find that on an average, 85% of users in the same cluster are likely to all exhibit good performance on our method, or all exhibit poor performance. This is an interesting result because we also find that users in the same cluster are four times more similar to each other in their credibility ratings than to users in other clusters. Although the similarity of ratings explains why our method performs similarly for similar users, an open question is whether the performance for a user goes up or down because of the cluster in which she is a member, or simply because the ratings given by her are too inconsistent to be captured by the Bayesian model.

As part of future work, we will try to identify more features to classify ratings, authorship, and social network matrices in terms of their characteristics to yield good or bad performance for users.

We also compared our method with other well known methods for trust and reputation computation meant for different applications. All these methods perform very close to random, even with personalization. We believe this to be due to a fundamental drawback of these methods: they try to form an objective assessment of credibility for users and messages, which is not appropriate for participatory media content.

- An Eigenvector computation on $\mathbf{A}_c^T.\mathbf{R}_r$ by leaving out the social network part (eqn. 1), is identical to the Eigentrust algorithm [135]. The best choice of parameters could only give a performance of MCC = -0.015 ($\sigma = 0.062$). Eigentrust has primarily been shown to work in P2P file sharing scenarios to detect malicious users that inject viruses or corrupted data into the network. However, the P2P context requires an objective assessment of the trustworthiness of a user, and does not allow for subjective differences, as desired for participatory media.

- An Eigenvector computation on the social network matrix (eqn. 2), personalized for each user, is identical to the Pagerank algorithm used to rank Internet web pages [132]. However, this too performs poorly with an MCC = 0.007 ($\sigma = 0.017$). This suggests that users are influenced not only by their own experiences, but also by the judgement of other users in their cluster, and by public opinion. Methods ignoring these factors may not perform well.

- The beta-reputation system [139] is used in e-commerce environments to detect good or bad buying and selling agents. It estimates the credibility of agents in an objective manner using a probabilistic model based on the beta probability density function. Only the public opinion is considered; ratings are filtered out if they are not in the majority amongst other ratings. It too does not perform well in the context of participatory media, giving an MCC = 0.064 ($\sigma = 0.062$).

Our conclusion is that approaches which subjectively model credibility, allowing users to be influenced in different ways by different sources, perform better than objective modeling approaches.

7.3 Hybrid algorithm

In this section, we propose a method to combine the non-collaborative algorithm for predicting message usefulness, and the collaborative algorithm for predicting credibility, into a single hybrid algorithm for message recommendations. Recall the following features of the two types of algorithms, and proposed extensions:

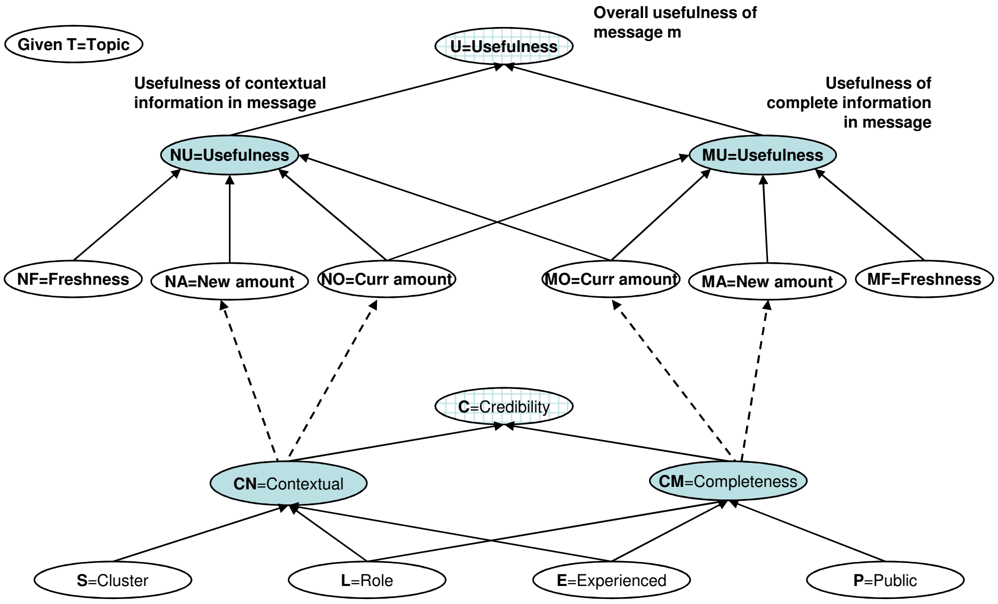


Figure 7.11: Hybrid model

- *Knowledge requirements*: Assumptions about knowledge of the social network of users, message content, and message authors, are common to both the non-collaborative and collaborative algorithms.

- *Non-adjacent clusters*: The current formulation of the non-collaborative algorithm is limited because it does not operate across non-adjacent social network clusters. The collaborative algorithm presents a feasible extension technique:

  - The collaborative algorithm can observe context and completeness similarities between pairs of users ($\mathbf{Y[n,n,4]}$) by correlating users on the **CN** and **CM** hidden variables. These similarity coefficients can be aggregated across users belonging to the same cluster, and hence detect context and completeness similarities between non-adjacent clusters as well. We refer to this as $\mathbf{Z}[n_c,n_c,\mathbf{4}]$, where n_c is the number of clusters.
  - These similarity coefficients can be used to extend the context and completeness models described in Chapter 6 in a straightforward manner. Virtual nodes can be created to emulate the presence of a user in any cluster, and the context and completeness estimates can be weighted on the context and completeness similarity between the non-adjacent virtual cluster and the original cluster of the user.

- *Ratings*: The non-collaborative algorithm requires usefulness ratings given by users to messages, and these ratings can remain local to each user. The collaborative algorithm requires credibility ratings given by users to messages, and these ratings need to be reported to a central agency. An apparent drawback in combining the two algorithms is the requirement for two sets of message ratings by the users. Due to a lack of suitable datasets, we can only speculate about possible solutions at this point:

  - Usefulness ratings can be estimated through implicit observations such as the time a user spends on a message, actions that a user takes after reading a message (for example, bookmarks the message, or appends a tag to the message), etc. Thus, the only requirement is for the user to give explicit credibility ratings to messages that can be reported to a central agency. This is a reasonable requirement to impose because ratings given by users on websites such as Digg are already publicly visible.
  - Alternatively, both the ratings can be considered to be equivalent with the hope that this will only result in additional noise in the learned model, which may get smoothed out as more and more data is collected over time.

- *Credibility assessment for usefulness model*: The credibility assessment method used in Section 7.1 was quite naive. This can be improved by instead using the credibility model of the collaborative algorithm to infer instances of the hidden variables of **CN** and **CM** for the contextual and completeness credibility of messages. These can then be plugged into the *na,no,ma,mo* variable estimates for the usefulness model as shown in Fig.7.11. The dashed lines indicate that the estimation of *na,no,ma,mo* uses the credibility values estimated by the credibility model.

Carrying forward these assumptions, we outline a hybrid algorithm, shown as Algorithm-7.4 and Algorithm-7.5. We leave the collection of a suitable dataset and evaluation of the combined algorithm to future work.

- **Learning** – Given the social network of users, messages, message authors, credibility ratings, and usefulness ratings:

  1. For each user n_i and topic t, learn the credibility models independently as in Section 7.2.
  2. Use the learned model parameters to compute context and completeness similarities between non-adjacent clusters ($\mathbf{Z}[n_c,n_c,\mathbf{4}]$) as proposed in this section.
  3. Extend the context and completeness estimates as proposed, and use the data to learn the usefulness models for each user as in Section 7.1. Rather than use the naive credibility assessment as in Section 7.1, use the contextual and completeness credibility estimates produced by the credibility model, as shown in Fig.7.11.

- **Inferring** – Given a new message m and recipient user n_i:

    1. Use the collaborative algorithm as a first-level filter to determine whether n_i will find the message to be credible.
    2. If so, infer the instance of **CN** and **CM** as the contextual and completeness credibility of the message. Use this to infer the usefulness of the message for the user, as shown in Fig.7.11, and decide if the message should be recommended to the user.

Although we are unable to evaluate the hybrid algorithm in this thesis, we hope that the strong foundations of our algorithms in sociological theory and information science will yield positive results, or at least encourage researchers to explore multi-disciplinary methods for message recommendation in greater detail.

Algorithm 7.4: Hybrid algorithm – learning

```
Scope: Given topic = $t \in T$
Input: $G_t \sim \mathbf{N[n,n]}$                         // Topic specific social network
       Message $M$ about event $e$ related to topic $t$
       $A(M) \sim \mathbf{A[k,n]}$, $B(M)$                 // Msg authors and content
       $M'_i \subseteq M$ read by user $i$, $U_i(M')$, $\mathbf{C[k,n]}$   // Usefulness and cred ratings
1. Cluster $G_t$ for tie classification
2. Learn credibility models
Compute $\mathbf{P[k]}$, $\mathbf{E[n,k]}$, $\mathbf{S[n,k]}$, $\mathbf{P'[k]}$, $\mathbf{E'[n,n]}$, $\mathbf{S'[n,n]}$
forall $i \in 1..n$ do
    Learn $\mathrm{P}_{it}(\mathbf{C} \mid \mathbf{E,S,P})$
3. Compute user–user and cluster–cluster similarity matrices
forall $i \in 1..n$ do
    Use learned credibility model to infer:
        $\mathbf{CN,CM} \mid \mathbf{E[i,}m \in M'_i\mathbf{]},\mathbf{C[i,}m \in M'_i\mathbf{]},\mathbf{P}[m \in M'_i]$
Compute $\mathbf{Y[n,n,4]}$
Compute $\mathbf{Z}[n_c,n_c,\mathbf{4}]$
4. Learn usefulness models
Compute required quantities such as $C_V,\gamma_i,\gamma_i^j$, using $\mathbf{Z}[n_c,n_c,\mathbf{4}]$
forall $i \in 1..n$ do
    Learn $\mathrm{P}_{it}(\mathrm{U}(M'_i) \mid \mathrm{A}(M'_i),\mathrm{B}(M'_i),\mathrm{C}(M'_i))$
```

7.4 Discussion and related work

Our work is different from most other traditional recommender systems such as those outlined in [119,120] because we develop models based on features of context and completeness examined by research in news media, which are not explicitly

Algorithm 7.5: Hybrid algorithm – make recommendation decision

```
Input: User $i$, Message $m$
      Quantities computed during the learning stage:
          $\mathbf{P[k]}$, $\mathbf{E[n,k]}$, $\mathbf{S[n,k]}$, $\mathbf{P'[k]}$, $\mathbf{E'[n,n]}$, $\mathbf{S'[n,n]}$
          $\mathbf{Y[n,n,4]}$, $C_V$, $\gamma_i$,$\gamma_i^j$
      Ratings $\mathbf{C[n,m]}$ given by other users to $m$
Output: Will user $i$ will find message $m$ to be useful?

Compute weighted mean of credibility of $m$ for $i$ over users similar to $i$
if weighted mean > recommendation threshold then
    Compute evidence variables $\mathbf{E[i,m]}$,$\mathbf{C[i,m]}$,$\mathbf{P[m]}$
    Infer P($\mathbf{CN}$ = $cn$| $\mathbf{E[i,m]}$,$\mathbf{C[i,m]}$,$\mathbf{P[m]}$),
          P($\mathbf{CM}$ = $cm$| $\mathbf{E[i,m]}$,$\mathbf{C[i,m]}$,$\mathbf{P[m]}$)
    Compute $na_{im}, no_{im}, nf_{im}, ma_{im}, mo_{im}, mf_{im}$ using $cn$,$cm$
    Infer $\mathrm{P}_{it}(u_m \mid na_{im}, no_{im}, nf_{im}, ma_{im}, mo_{im}, mf_{im})$
    Make recommendation decision based on probability of usefulness:
          Depends on the relative true-positive and false-positive rates
          that user $i$ may be comfortable with
else
    Reject
```

modeled by other methods. Furthermore, the traditional approaches do not use information about the underlying social network of users.

More closely related work includes [118, 125, 126]. [118] makes recommendations based on stochastic simulations that replicate the observed patterns of information flow on social networks. [125] operates in a P2P setting, and uses decentralized CF algorithms executed within local social network neighborhoods of users. [126] learns content-based gradients on links between users; this can be used to route messages along desired gradients to users who will be interested in these messages. However, unlike our method which is based on the real-world social network of users, all these methods consider an artificial social network that is formed by linking together users observed to be similar to each other. Furthermore, these methods do not explicitly model message features such as context and completeness.

Our credibility modeling methodology is also different from most prior work in the area of trust and reputation systems. Various researchers in the P2P community have focused on Eigenvector based methods to compute the reputation of peers in sharing reliable content [132, 135]. The ratio of successful to unsuccessful content exchanges is computed for each pair of peers who have interacted in the past, and these values are propagated in a distributed manner assuming a transitive trust relationship between peers. However, this is used to only compute the peer reputations (i.e. evaluating users) and not the reliability of content that is shared by the peers. A similar approach of Eigenvector propagation was also used in [136] to compute reputation scores in a blog network, but the reputation of individual blog-entries was not computed. In our approach, we make use of message ratings and compute the credibility of each message.

For P2P networks, a method was proposed in [137] where the object reputation is directly calculated to determine whether to accept a file being shared on a peer-to-peer network. Transaction history is used to assign edge weights between pairs of peers based on the similarity of ratings given by them to common objects rated in the past. Instead of using Eigenvector propagation to compute an absolute reputation score, a small set of shortest paths is found for each pair of peers, and the relative trust between the peers is computed as the mean of the product of edge weights along the paths. In our approach, we offer a richer multi-dimensional representation, integrating concepts of cluster, experienced, and public credibility.

Researchers in the AI community have examined trust models for multi-agent based electronic marketplaces. For example, [138] and [139] offer systems that determine the trustworthiness of an agent (i.e. a user). In addition, the use of an extensive trust model is promoted in [103], to include features of contextual, role-based and experienced trust. We also have a multi-dimensional model, but we place great emphasis on representing and making use of the social network of a user, in order to learn a user-specific credibility rating for messages. This gives us improved performance over other methods [193, 194]. For future work, we plan to extend our algorithms in a number of ways.

Confidence bounds: Methods for combining trust and confidence have been proposed by researchers such as [140, 141]. It may be valuable to explore how to incorporate the concept of confidence into our models, for example as a way of placing bounds on the statistical hypotheses that are formed at each step of our algorithms.

Model extensions: We view our proposed method more as an extensible framework that can be extended to incorporate new insights or information. For example, we could explore the concept of *expert credibility* in the future, for which we would repeat the Eigenvector computations by considering ratings only by a specific set of users categorized as expert users by expert identification algorithms [142]. Another important piece of information that is typically available in participatory media content, although it is not available in the Digg and Orkut datasets that we used, is the *message link matrix* based on hyperlinks between messages. A rule that credible messages link to other credible messages can be modeled through pagerank or HITS, and included as an additional weighting factor in the Eigenvector computations. Alternatively, the polarity between links can be derived by sentiment analysis of the anchor text around a hyperlink [124], and distrust propagation methods can be used to produce credibility scores based on the message link matrix [143].

Robustness to attacks: It would be desirable to have our model be robust in the face of attacks by malicious users. This may include scenarios where attackers could add noise to the ratings matrix by giving random ratings to various messages, or attackers could pollute the social network matrix by inviting unsuspecting users to link to them as friends, or even more sophisticated scenarios where attackers could collude with each other. In future work, we would like to examine the robustness of our model against such types of attacks. We also believe that attack analysis could give important insights about the implicit interactions between various pieces of information that are modeled together; such insights are likely to be valuable to improve the performance further.

Optimized computation: The proposed models may be computationally intensive. However, Eigenvector optimization schemes are available that can decompose a large matrix into smaller matrices, and then combine the components together in an approximate fashion [144]. We will experiment with such schemes in future work. We also outline a different aspect of scalability related to reducing the volume of message updates in the next chapter.

Exploration-exploitation tradeoff: We hinted in Table 5.2 that the recommendation algorithm we have outlined so far aims to replicate the observed user behavior. However, the algorithm should also be capable of helping the user to explore new territories – topics, opinions, etc – if the user so desires. This is referred to as the *exploration-exploitation tradeoff* [145], and we will discuss such tradeoff management in the next chapter as well.

Chapter 8

Recommender system architecture

Caught in the vortex of publicity that is staged for shorn or manipulation, the public of non-organized private people is laid claim to not by public communication but by the communication of publicly manifested opinions –
Jurgen Habermas, The Structural Transformation of the Public Sphere, 1962

In the previous chapter, we described variants of recommendation algorithms based on learning the *usefulness* and *credibility* models for users in a personalized manner. We assumed that all the required knowledge and computation for learning the models was available centrally. However, scalability of computing training sets and learning the models for hundreds of millions of users can be quite challenging to ensure centrally. In this chapter, we delve deeper into the system design and separation of computation roles into different components to ensure scalability. Although we cannot implement such a system as part of this thesis, in Section 8.1 we outline the system architecture and identify open research and engineering problems that will need to be solved in order to build the system. In Section 8.2, we then focus on one component of the system related to the scalability of message updates, and propose a solution. Finally, in Section 8.3, we contrast our approach with related work.

8.1 System architecture

We envision our system to be similar to commercial blog-aggregators such as Technorati or Bloglines. These aggregators rely on blog hosting software to send updates about new blog posts written by users. The posts are then indexed, and the most popular posts, latest posts, posts on blogs marked as favorites by users, etc, are listed. In our system, we wish to receive similar updates, but rather than produce a simple listing of posts published on pre-selected blogs, we wish to run our personalized recommendation algorithms to infer those posts that would be of most

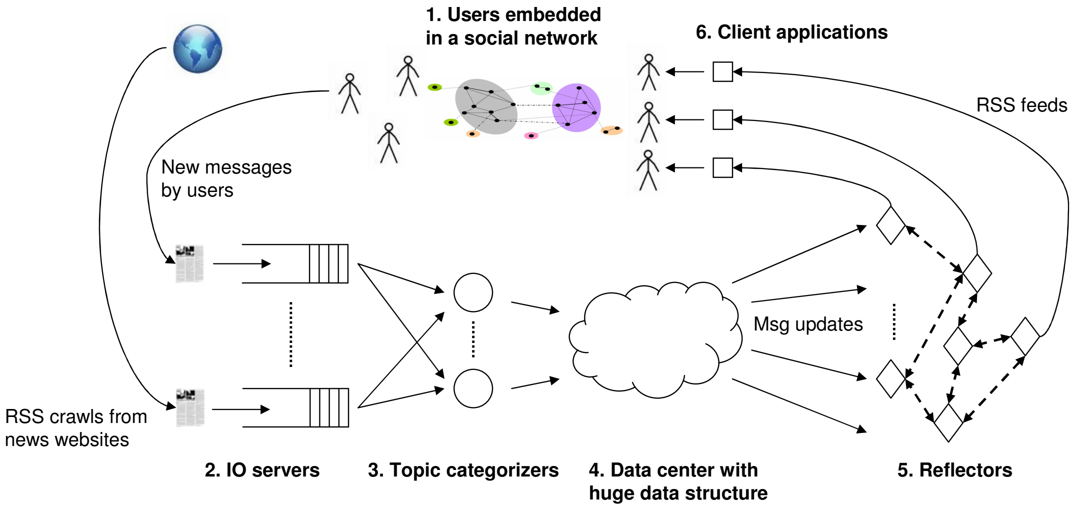


Figure 8.1: System architecture

interest to users. A scalable design for a blog and RSS-feed aggregation system with similar objectives, Cobra, was proposed in [146]. Cobra ensures scalability by partitioning functionalities such as content aggregation, topic categorization, and content indexing into multiple servers in a cluster. However, Cobra relies on users to specify keywords for blog posts that they would find interesting. Our system is more complicated because the recommendation algorithms make use of social network information, ratings, and past history of the users to produce recommendations. We therefore use the Cobra architecture as a starting point for our system design, and build upon it to accommodate our specific requirements. Fig.8.1 shows the proposed system architecture and the human environment in which it is intended to operate. The various components and their tasks are described below with reference to the hybrid algorithm given in the previous chapter.

1. *Data sources*: These are news or blogging websites that host content. They typically provide RSS feeds, or can be customized to "ping" aggregation services, such as our system, whenever new content is published.

2. *I/O servers*: These are responsible to download new content from the data sources.

3. *Topic categorizers*: The downloaded content needs to be analyzed to categorize it into various topics. Post categorization, metadata about new messages, including information about the message author and topic, is pushed to the data center described next.

4. *Distributed data structure*: The primary task of the data center is to execute offline computations required to make recommendation decisions. This includes clustering of the social network graph (Algorithm-8.1.1), computation of public and cluster credibilities matrices (Algorithm-8.1.2), similarity

Algorithm 8.1: Hybrid algorithm – learning

```
Scope: Given topic = $t \in T$
Input: $G_t \sim \mathbf{N[n,n]}$                                              // Data center
        Message $M$ about event $e$ related to topic $t$                       // Data center
        $A(M) \sim \mathbf{A[k,n]}$, $B(M)$                                    // Data center
        $M'_i \subseteq M$ read by user $i$, $U_i(M')$                          // Client applications
        $\mathbf{C[k,n]}$                                                      // Data center

1. Cluster $G_t$ for tie classification                                        // Data center

2. Learn credibility models

Compute $\mathbf{P[k]}$, $\mathbf{E[n,k]}$, $\mathbf{S[n,k]}$, $\mathbf{P'[k]}$, $\mathbf{E'[n,n]}$, $\mathbf{S'[n,n]}$        // Data center
forall $i \in 1..n$ do
    Learn $\mathrm{P}_{it}(\mathbf{C} \mid \mathbf{E,S,P})$                    // Client applications

3. Compute user–user and cluster–cluster similarity matrices

forall $i \in 1..n$ do
    Use learned credibility model to infer:                                    // Client applications
        $\mathbf{CN,CM} \mid \mathbf{E[i,}m \in M'_i\mathbf{],C[i,}m \in M'_i\mathbf{],P[}m \in M'_i]$
Compute $\mathbf{Y[n,n,4]}$                                                    // Data center
Compute $\mathbf{Z}[n_c,n_c,\mathbf{4}]$                                       // Data center

4. Learn usefulness models

Compute required quantities such as $C_V$,$\gamma_i$,$\gamma_i^j$, using $\mathbf{Z}[n_c,n_c,\mathbf{4}]$    // Data center
forall $i \in 1..n$ do
    Learn $\mathrm{P}_{it}(\mathrm{U}(M'_i) \mid \mathrm{A}(M'_i),\mathrm{B}(M'_i),\mathrm{C}(M'_i))$          // Client applications
```

Algorithm 8.2: Hybrid algorithm – make recommendation decision

```
Input: User $i$, Message $m$
        Quantities computed during the learning stage:
            $\mathbf{P[k]}$, $\mathbf{E[n,k]}$, $\mathbf{S[n,k]}$, $\mathbf{P'[k]}$, $\mathbf{E'[n,n]}$, $\mathbf{S'[n,n]}$      // Data center
            $\mathbf{Y[n,n,4]}$, $C_V$, $\gamma_i$,$\gamma_i^j$                  // Data center
        Ratings $\mathbf{C[n,m]}$ given by other users to $m$                   // Data center
Output: Will user $i$ will find message $m$ to be useful?

Compute weighted mean of credibility of $m$ for $i$ over users similar to $i$    // Reflect
if $\textit{weighted mean} > \textit{recommendation threshold}$ then
    Compute evidence variables $\mathbf{E[i,m],C[i,m],P[m]}$                    // Client applications
    Infer $\mathrm{P}(\mathbf{CN} = cn \mid \mathbf{E[i,m],C[i,m],P[m]})$,       // Client applications
          $\mathrm{P}(\mathbf{CM} = cm \mid \mathbf{E[i,m],C[i,m],P[m]})$
    Compute $na_{im}, no_{im}, nf_{im}, ma_{im}, mo_{im}, mf_{im}$ using $cn$,$cm$       // Client app
    Infer $\mathrm{P}_{it}(u_m \mid na_{im}, no_{im}, nf_{im}, ma_{im}, mo_{im}, mf_{im})$      // Client app
    Make recommendation decision based on probability of usefulness:            // Client
          Depends on the relative true-positive and false-positive rates
          that user $i$ may be comfortable with
else
    Reject
```

matrices between pairs of users (Algorithm-8.1.3), and the clustering and integration coefficients (Algorithm-8.1.4). Consequently, the data center has to use specialized data structures to store the required data which includes the topic specific social network graphs, messages in circulation in the system, their participants and various contributions, credibility ratings, and similarity matrices. Note that the data centers do not actually execute the model learning steps shown in Algorithms 8.1.2 and 8.1.4, but only compute the necessary data required to learn the models. This data is appended to the message metadata, and eventually makes its way to the client applications via reflector nodes.

6. *Client applications*: Client applications run on user devices such as laptops and mobile phones. These applications receive message updates from the reflector nodes (described next), and use message metadata and ratings given by the user to learn the usefulness and credibility models for the user (Algorithms 1.2 and 1.4). For new messages, message metadata is used to draw inferences from the usefulness and credibility models to decide which messages should be pushed to the user (Algorithm-8.2). This design improves both scalability and privacy. The Bayesian models are learned in a decentralized fashion on the client side, saving valuable computation cycles for a centralized approach. These client-side learning algorithms themselves have low computation costs, as seen by us in our experiments in Chapter 7. Privacy is also improved because usefulness ratings need not have to be conveyed to a central agency. This makes it simpler to closely monitor user activities such as the time spent by a user on a message, clicks generated by the user, etc.

5. *Reflectors*: Reflectors are client nodes that are elected as super nodes, or are hosted on servers distributed across the world. A reflector instance exists for each cluster in a topic specific social network, and hosts an RSS feed for messages likely to be relevant to users of that cluster. These feeds also carry the message metadata required by the client applications. Reflectors essentially perform two functions:

   - Whenever a new message enters the system, the message and its associated metadata have to be pushed out to various client applications. This can cause significant network overhead in message exchanges if a central message registry is maintained in the data center. Instead, reflectors now perform the same function – message states are distributed across multiple reflectors rather than being hosted centrally.
   - Even if the message states are distributed across multiple reflectors, we need to ensure that no single reflector is overloaded with too much message state. In addition, since the client applications make the final recommendation decision, a pre-filter step is needed so that client applications are not overloaded with too much message traffic from which

to make recommendation decisions. Both these objectives are achieved by connecting the reflector nodes in a certain overlay topology, and running routing and forwarding algorithms on this topology which implicitly perform the pre-filter step by propagating messages only to candidate reflector nodes that may find the messages to be relevant. This is essentially the *if then* step shown in Algorithm-8.2, to run the decision step of the collaborative-based algorithm in a distributed fashion.

Thus, whenever information about a new message is received and updated in the data center, a few reflectors are notified about the message. The initial choice of reflectors may include those corresponding to the cluster of the message author, and the clusters adjacent to that. Subsequently, based on ratings given to the message, routing algorithms operating between reflectors spread message updates to other relevant reflectors. This helps to propagate updates about new messages to appropriate destination reflectors without causing significant network overhead in push or pull of message updates from a central message registry. This design has another advantage: RSS feeds hosted by a reflector are downloaded periodically in batches by client applications in the same cluster as the reflector node. Therefore, continuous connectivity with the Internet is not required by the client application, making the applications ideal to run in a delay-tolerant fashion on mobile phones using OCMP, or even in remote rural stations using KioskNet.

A distributed system naturally implies the need for protocols to transmit required information to various entities in the system. Furthermore, many parameters such as the number of servers in the data center, bandwidth requirements, etc, will have to be intelligently decided. Clearly, a significant portion of system building will be a challenging engineering effort. But interesting research problems need to be solved for each of the components, and are described next.

Open problems

- **Geographic locationing**: The I/O servers are likely to be bandwidth bound; hence, they should be geographically distributed to download content from sources close to them. Network location services such as Vivaldi [155] may be suitable to find an optimal assignment of I/O servers to data sources.

- **Identity management**: The same user may subscribe to many different blogging and social networking websites, each of which have their own userids. This will make it hard to maintain a single identity for each person. The issue should hopefully get resolved through consortiums such as the OpenSocial initiative (*code.google.com/apis/opensocial/*), but techniques for entity resolution to match identities and aliases may be required otherwise [156].

- **Topic categorization**: Since our models are based on topic specific social networks, we have to classify new messages into their appropriate topics. In some cases such as the resources we used for our experiments, content is published in specific channels which makes topic inference trivial. In other cases, the topics will have to be inferred from the content itself. Fast indexing techniques such as those used in publish-subscribe systems [149] may prove useful to match tags or keywords with pre-defined topics, or document classification techniques such as Latent Semantic Indexing [158] can be used to identify the topics. Similarly, the interests of a user in various topics may be determined from the membership of users in discussion groups [191].

- **Efficient data structures**: The design of efficient data structures to store large social network graphs, such that clustering and computation operations on these graphs can be executed efficiently, is an interesting research problem. Relevant techniques include the design of structure indices [150] using virtual coordinates and landmark methods, that can be used to optimize the computation of metrics such as shortest paths, edge betweenness, and centralities in large social network graphs.

- **Reflector topology**: To enable efficient propagation of message updates from the data center to client applications, an overlay topology connecting the reflectors needs to be developed with suitable routing algorithms operating on this overlay. We show in the next section that it is possible to construct topologies such that even naive routing algorithms operating on this overlay will not pose a significant demand on the size of the routing state that needs to be maintained at each node, similar to the work described in [154]. Manageable routing state sizes generally indicates a scalable routing infrastructure, even though it does not say anything about the possibility of breakdowns from traffic overload at the reflector nodes or client applications. A good solution to the routing and forwarding problem described next, is required to ensure safety from traffic overload.

- **Reflector routing and forwarding**: The routing algorithms operating on the reflector topology should be capable of routing messages to relevant destination nodes that would find the messages to be worth recommending to their users. We do not solve the routing and forwarding problem in this thesis, but outline a possible approach as follows. Given knowledge of $\mathbf{Z}[n_c,n_c,\mathbf{4}]$, clusters similar to each other can be identified using dimensionality-reduction techniques such as **Principal Component Analysis (PCA)**. Subsequently, directed-diffusion [151] or content-based routing algorithms [157] can be used to establish routing tables on the reflector overlay to enable reflectors of similar clusters to route messages to each other. In this manner, the decision step for the recommendation becomes distributed, because only messages from relevant clusters will make their way to a destination reflector. This reflector node can then run an additional check based on ratings given to messages by users in these clusters, to eliminate all but a likely candidate set of messages.

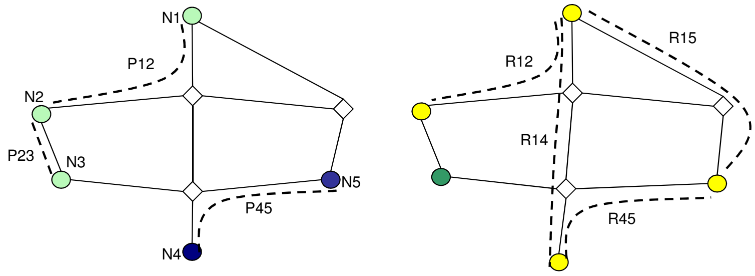


Figure 8.2: Example of a reflector graph

> The metadata for these messages can then be hosted in the RSS feed so that client applications associated with the reflector node can receive the updates. Alternatively, in case message traffic is very heavy for destination reflector nodes to perform the filter check, more intelligent forwarding algorithms will be needed which perform active filtering while messages are in transit, and preemptively drop messages unlikely to be relevant for particular reflectors.

In this thesis, we only focus on the problem of overlay topology construction for reflectors. This does not mean that the other problems are not as significant, but only that we choose this specific problem because it remains an unexplored area and we already have relevant datasets for evaluation. We next describe the problem in more detail and present results of our proposed solution.

8.2 Reflector topology

We begin by describing reflectors in more detail. It was shown in the last chapter how the similarity matrix $\mathbf{Y}_t[\mathbf{n},\mathbf{n},\mathbf{4}]$ for topic t could be constructed to capture four kind of similarities between pairs of users – contextual similarity, completeness similarity, and cross similarities between context and completeness. We then explained that user–user similarity could be aggregated into cluster–cluster similarity by averaging the similarities across all pairs of users in each cluster pair. This gives the matrix $\mathbf{Z}_t[\mathbf{n}_c,\mathbf{n}_c,\mathbf{4}]$, where n_c is the number of clusters in the topic specific social network for t. PCA on each of $\mathbf{Z}_t[\mathbf{n}_c,\mathbf{n}_c,\mathbf{0}]$, $\mathbf{Z}_t[\mathbf{n}_c,\mathbf{n}_c,\mathbf{1}]$, $\mathbf{Z}_t[\mathbf{n}_c,\mathbf{n}_c,\mathbf{2}]$, and $\mathbf{Z}_t[\mathbf{n}_c,\mathbf{n}_c,\mathbf{3}]$ would yield low dimensional identification vectors for each cluster, such that clusters similar to each other would also be similar on their identification vectors. Since each cluster has a corresponding reflector node, for ease of exposition, we consider each node as having a colour, and nodes of similar clusters as having the same colour.

Fig.8.2 shows hypothetical graphs of reflector nodes. Two topics are shown, represented as circles and diamonds respectively. The left part of the figure shows

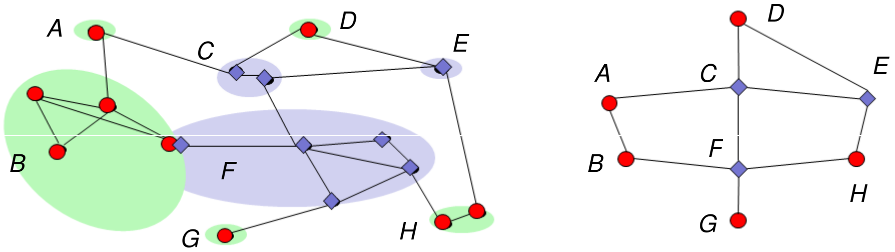


Figure 8.3: Reflector topology construction

colour assignments to nodes denoting contextual similarity. Nodes having a similar context (or colour) can route messages to each other along routes shown as {P_{12}, P_{23}, P_{45}}. The right part of the figure shows colour assignments for completeness similarity, with routes between completeness-providing clusters shown as {R_{12}, R_{14}, R_{15}, R_{45}}. Similarly, two additional colour distributions will exist for the remaining two types of similarities.

Given this representation, routing algorithms are needed which can find efficient routes between nodes of the same colour. We show that we can construct a reflector node graph such that even naive routing algorithms will not impact the scalability of message updates on the reflector graph. We next describe our topology construction method, and then present results from analysis of the Orkut dataset described in Chapter 6. As we explain later, our topology construction method exploits the properties of *navigability* and *topic locality* in social networks.

8.2.1 Topology construction

We add an edge between two reflector nodes if their respective clusters have a user in common; or there is a pair of users, one in each cluster, having a social network tie between them. This represents the intuition that users are aware of their own topic interests (of course) and the topic interests of their friends, so that an edge is added in the reflector graph to connect nodes having common users or friends respectively. Fig.8.3 shows a social network on the left that would result in the reflector topology on the right. Labels A, B, etc represent clusters, which get assigned to the corresponding reflector nodes. All users in the social network are shown to be interested in only a single topic, except for the user present along the edge of clusters B and F which results in a link between the reflector nodes for B and F.

We are primarily motivated to try this approach because of the *navigability property* of social networks. It has been noticed in multiple experiments [152, 153] that people are able to route messages to a specific person along friendship chains by only making use of information about their own immediate friends, such as the profession, geographical location, and interests of friends. For example, a source who wants to send a letter to a stockbroker in Boston but does not know the address of the broker, would send the letter to a friend who probably knows somebody in Boston, or is interested in finance. This friend would then forward the letter to another friend who probably lives in Boston, etc. This example actually comes from an experiment done by Milgram in 1958 [152], illustrating the navigability property of social networks.

Watts et al proposed a multi-dimensional model for social networks to explain the navigability property [154]. In this model, each person is identified by a vector representing features such as the profession of the person, geographical location, interests, etc. Under the homophily assumption that people are likely to associate with other people having a similar identity vector, and the additional assumption that each person knows the identity vectors of her social network neighbors, it was shown that with a high probability messages could be routed to a desired destination using a simple greedy algorithm. Not only could routes be found in this model, but stringent bounds on the route length could also be established. Since this model only made use of the local routing state of size O(degree of node) per node, we were encouraged that it may indeed be feasible to find efficient routes also on the reflector graph which is constructed on the same principles (here, each colour per topic is considered as an independent dimension). Of course, generalization of the insights of the Watts et al [154] model is not straightforward:

- The reflector graph is at the cluster level, while the model operates at the user level.
- The model assumes the same distribution of identity across all dimensions, whereas some topics can be more popular than others in the reflector graph.
- The model works only for a certain space of homophily and number of dimensions, with the number of dimensions typically having a small value ~ 10 for $\sim 10^5$ people. In fact, for the greedy routing algorithm to succeed, the number of dimensions should decrease with an increase in the size of the social network so that the per-dimension graphs never become partitioned. However, the topic node graph may have hundreds of topics, with many colours ($\sim$ dimensions) within each topic, and the number of dimensions is likely to increase with an increase in the number of users. Thus, greedy routing algorithms operating on the homophily assumption may altogether fail to find routes across partitions. We would therefore like to develop algorithms which work despite these constraints, potentially by making use of routing state greater than O(degree of node) per node in the reflector graph.

However, the model served its purpose by providing us the insight that navigability using local routing state may be enabled by the inherent homophily properties of social networks. Therefore, we next analyze the structure of the network using the Orkut dataset. Our goal is to examine whether it is even feasible and practical to build routing algorithms on this reflector graph. We do recognize that our analysis has limited generalizability because of an evaluation on a single dataset from Orkut; we will do more extensive evaluations in future work.

We analyze three subgraphs of our original Orkut dataset of 42000 users, containing approximately 10000, 17000, and 23000 users each, so that we can generalize our conclusions independent of the number of users. These datasets contain 92, 165, and 301 topics respectively. We first describe the characteristics of these datasets at the user level, and later analyze the reflector graph at the cluster level.

Note that since we do not have any knowledge of the colour distribution over the topic node graph, we only determine the connectivity properties between nodes of the same topic. However, we consider the extreme case that every node must be able to find a route to every other node in its topic. This represents the worst case when all nodes are of the same colour, and hence interested in each other's messages.

8.2.2 Characteristics of datasets

The datasets were obtained in three stages. We started with a randomly chosen strongly connected cluster of 200 users from our original Orkut dataset. Then we expanded this graph to include all friends of these users, and enumerated topics of interest to these users. This gave us the first dataset of 10000 users and 92 topics. For the second dataset, we included friends-of-friends of some users by localized expansion of the first dataset graph in the direction of a few randomly chosen dense clusters in the graph. This gave us the second dataset of 17000 users and 165 topics. For the third dataset, we similarly expanded the graph by including more friends-of-friends in other directions of the graph. This gave us the third dataset of 23000 users and 301 topics.

Interestingly, the number of topics of interest grows almost linearly with the number of users. This can be explained by the existence of topic locality, that topics of interest tend to be localized in social network neighborhoods of users close to each other, much like topic locality noticed in P2P networks [159]. Locality is natural to arise in the presence of homophily in user interests and size restrictions on the number of users per topic. We notice both properties in our datasets; subsequent experiments explained next give further evidence for the locality hypothesis.

Fig.8.4 shows the degree of homophily, calculated as the fraction of links of a user to other users having at least one common topic of interest. Interestingly, this is an almost uniform distribution across the datasets. Considering that less than half the users in each dataset have homophily = 0, this gives evidence of the existence of homophily in the network.

All datasets also exhibit a power law in topic popularity between topic sizes of 100 to 1000, truncated below and above these thresholds. The power law distribution sets our dataset different from the dataset used for validating the Watts et al

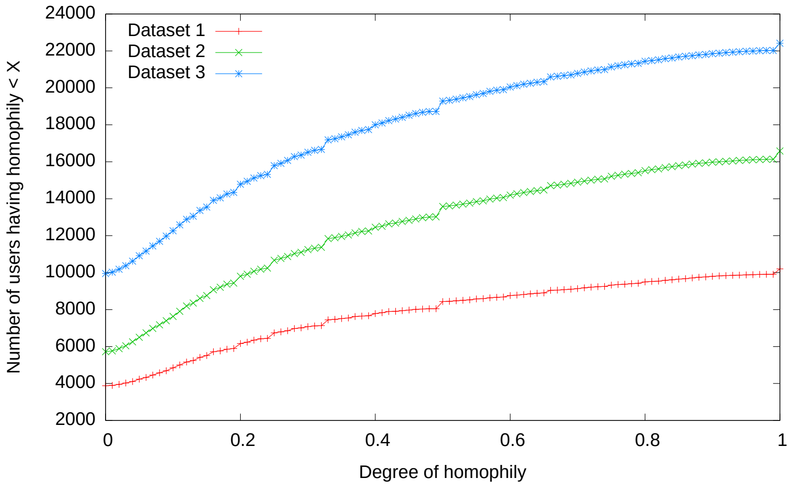


Figure 8.4: User homophily

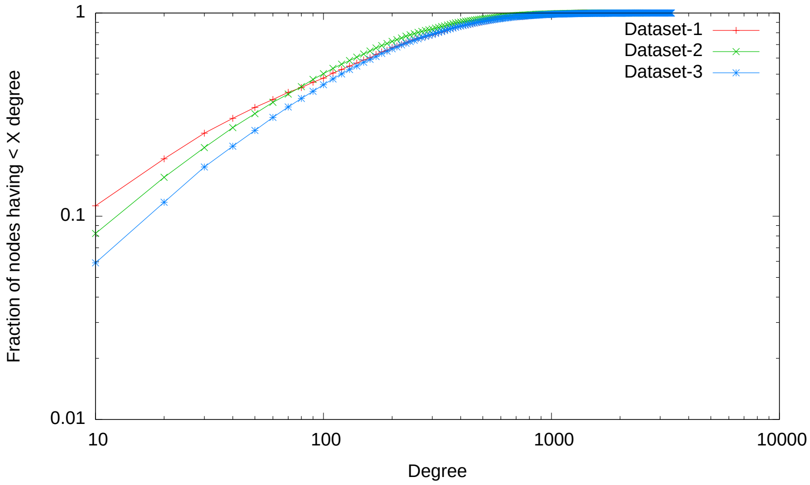


Figure 8.5: Degree distribution of reflector nodes

model [154]. Through a separate analysis, we also found that 95% of the users in the datasets were interested in up to 10 topics.

8.2.3 Reflector graph analysis

We next construct the reflector graphs for the datasets using the clustering methodology described in Section 6.3.2. Table 8.1 describes some properties of these graphs, and Fig.8.5 shows the degree distributions. The average degree is 75 per node, indicating a considerably dense graph. As would be expected, the degree distribution of the reflector graph is dependent only upon the structural properties of the social network of users and the user-topic distributions, but independent of the number of users included in the datasets. This is encouraging because it means that even if the social network graph is expanded to include more and more users, and hence more

Table 8.1: Characteristics of datasets

Dataset	1	2	3
# of users	10514	17108	23025
# of topics	92	165	301
# of nodes	1899	4580	6593
Max diameter	5	6	7
Avg diameter	2.31	2.50	2.52

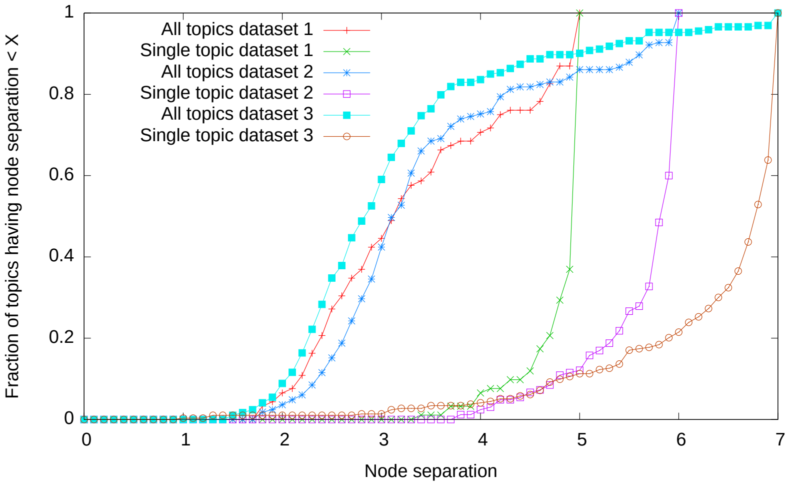


Figure 8.6: Reflector node separation

and more topics, the degree distribution characteristics of the topic node graph will not change, and hence properties of the routing algorithms dependent upon the degree distribution will also remain stable.

It is possible that routing algorithms could establish most routes on these graphs as passing through high-degree nodes. Since we do not want to overload any node with a heavy amount of routing state, we tried removing all nodes with degree greater than 400. This did not break the connectivity of the graphs or alter the maximum and average diameters. Therefore, we do not consider these nodes for estimation of the routing state in further analysis of the datasets. Note that this causes us to over-estimate the bounds on the routing state, but as shown later, the results turn out to be encouraging despite that.

Fig.8.6 shows the distribution of average length of shortest paths between nodes of the same topic in the different datasets. The CDFs to the right consider each topic specific network individually, which can be sparse and may even contain disjoint subgraphs. The separation between nodes that did not have a path to each other was assumed to be the maximum diameter in the respective dataset. The CDFs to the left consider the entire graph which allows paths to hop across different topics. As expected, the node separation improves considerably; the average separation is ~ 3, with 80% of the topics having a separation less than 4. This

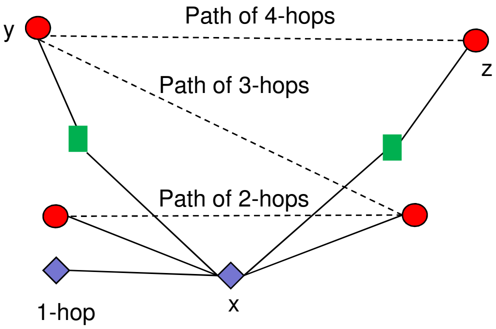


Figure 8.7: Number of routes passing through a node

is encouraging because it shows that short routes exist between nodes of the same topic even if the routes have to pass through reflectors that do not carry feeds for that topic.

What is more interesting, however, is that the distributions remain roughly consistent across the three datasets, indicating that node separation is independent of the number of users. A possible explanation is that of topic locality. This is also encouraging because it means that nodes need not flood their interests very far into the reflector graph, and hence not increase the size of the routing tables of other nodes.

We finally compute the approximate upper and lower bounds of routing state to be maintained at each node. As explained earlier, we do this for the worst-case where all nodes of a topic are assumed to be of the same colour, and hence messages should be routable between all pairs of nodes of the topic. Further, we assume a simple routing algorithm that establishes at least one route for every pair of nodes per topic. Thus, rather than establish a single route $\{a$–b–$c\}$ for nodes $\{a, b, c\}$ of the same topic, the algorithm may establish overlapping routes $\{a$–b–c, b–c, a–$b\}$, each of which is maintained separately in the routing tables of intermediate nodes. Therefore, we over-estimate the routing state that will be required.

We find the upper bound by first computing the shortest paths d_{ij} between all pairs for nodes. Then, for each node x, we compute the number of pairs of 1-hop, 2-hop, etc neighbors (y, z) of the same topic such that the shortest path lengths between the neighbors $d_{yz} = d_{yx} + d_{xz}$. This is shown in Fig.8.7 – the upper bound of routing state maintained at a node is essentially a count of the number of shortest paths that run through the node. In other words, this gives a worst case bound of the maximum number of shortest paths a routing algorithm may establish between nodes of the same topic (y, z) that can possibly pass through some other given node (x). For the lower bound, we include only a single path between (y, z) in the count across all nodes; that is, for a given ordering of choosing nodes x_1, x_2..., we count (y, z) only the first time it occurs as a shortest path passing through some node x_i. Thus, for any pair of nodes, a path between them is counted only once.

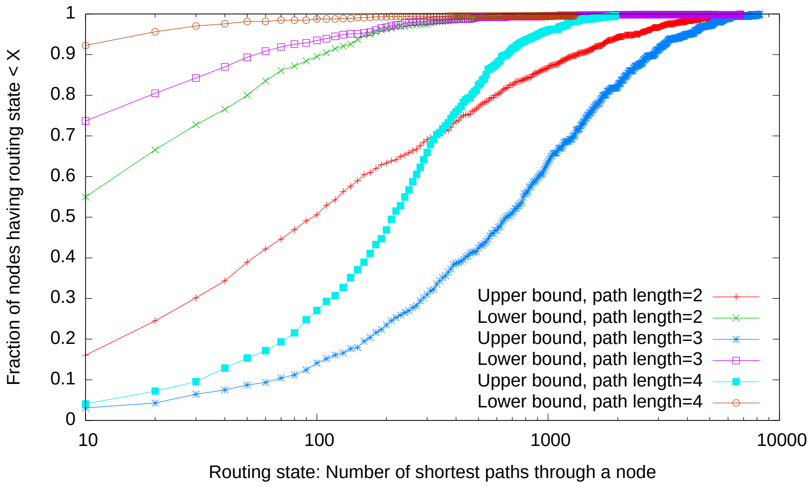


Figure 8.8: Upper and lower bounds on dataset-1: Number of users = 10514

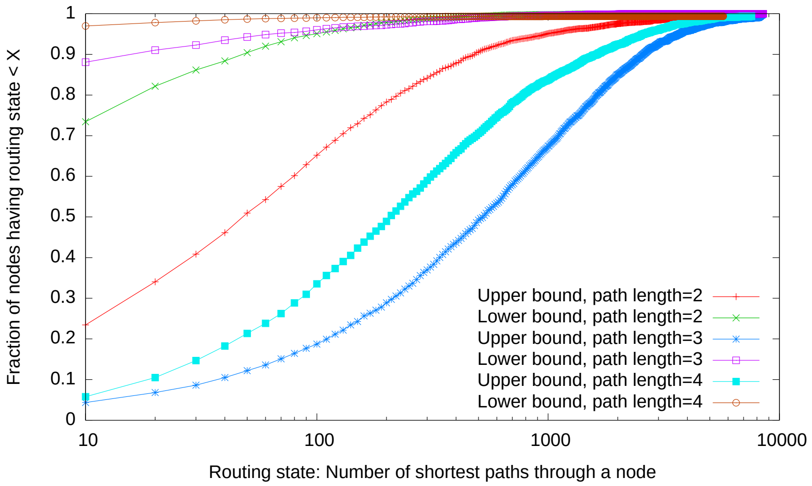


Figure 8.9: Upper and lower bounds on dataset-2: Number of users = 17108

After repeated iterations with different orderings of choosing nodes x_1, x_2..., the mean distribution of the number of shortest paths across different orderings gives the lower bound distribution.

Fig.8.8, 8.9, and 8.10 show the upper and lower bounds for the routing states in the datasets. Only the amount of state for $d = \{2,3,4\}$ is shown, 3 being the average separation between nodes of the same topic, and 4 being the 80%ile separation as shown in Fig.8.6. The amount of state for $d = \{1,5,6\}$ was insignificant. It can be seen that 90% of the nodes need state < 4000. The bounds generally increase from d=2 to d={3,4} but the lower bounds remain below 500, while the upper bounds remain below 4000. This indicates a total routing state of size < 10000 in the worst case for graphs with an average node degree of 75 and average separation of nodes of the same topic as 3. Comparing Fig.8.8, 8.9, and 8.10, the distribution of upper bound routing state remains nearly independent of the number of users, giving

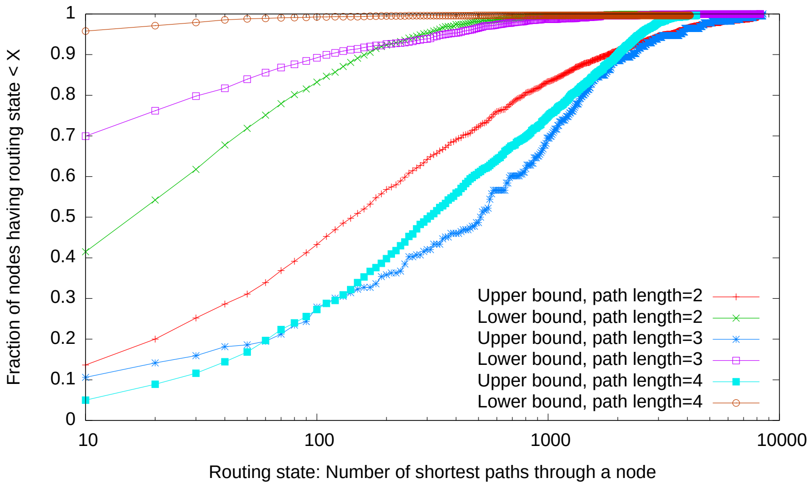


Figure 8.10: Upper and lower bounds on dataset-3: Number of users = 23025

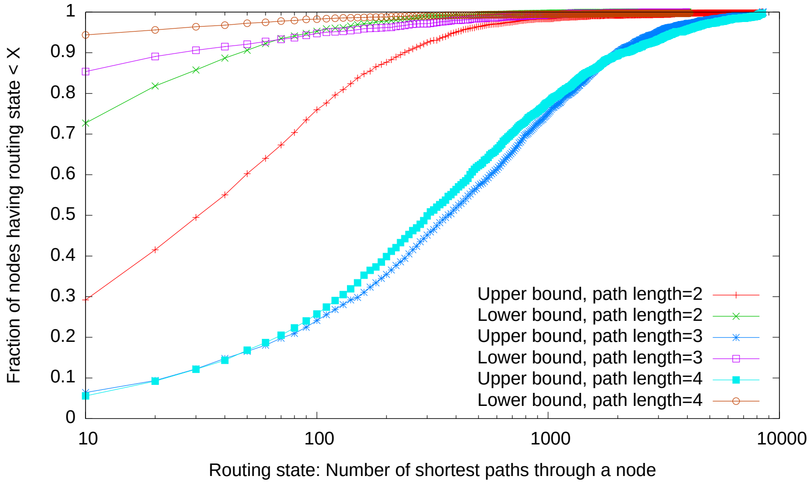


Figure 8.11: Upper and lower bounds considering only popular topics

more evidence of topic locality in the network. This is not unexpected because the routing state at a node depends upon the distribution of nodes of various topics in the neighborhood, and hence upon the node degree (Fig.8.5) and node separation (8.6) distributions. And these have already been seen to be independent of the number of users.

We add another degree of parametrization at this stage, by analyzing the topic node graphs separately for two cases: including nodes of all topics, and nodes of only popular topics. We characterize unpopular topics as those having less than 200 users, which represents a 20% cutoff mark from the maximum of approximately 1000 users in very popular topics.

Fig.8.11 shows that for the second dataset, the bounds for routing state are consistent at the nodes, when considering only popular topics of more than 200 users. Compared with Fig.8.9 which includes all topics, it is evident that the inclusion of unpopular topics does not alter the routing state significantly, and in fact, improves the routing state in many cases. This is promising because it indicates that unpopular topics are distributed across the graph in such a way that they do not cause a significant increase in the routing state. Therefore, at least in this dataset we can rely on popular topics to ensure connectivity, while unpopular topics usually “hang” off the popular topics. In future work, we will develop mathematical models to gather insights about the underlying reasons that may give rise to such characteristics, as well as study other datasets to determine if this property holds in other situations.

8.3 Discussion and related work

Most work in the area of blog-aggregation and publish-subscribe systems seems to have focused on scalability aspects such as the crawling and polling of web pages, rather than on the design of systems that can run personalized recommendation algorithms based on social networks and user ratings. For example, a distributed publish-subscribe system, Corona [147], updates users by sending instant messages about new publications on pre-specified URLs, unlike our requirement of producing automated recommendations. Another system, Herald [148], relies on publishers and subscribers to agree on a set of topics published in different channels to which users subscribe, but again, does not provide any recommendation services. The Cobra [146] blog-aggregation system does provide personalized RSS feeds, but is designed around a keyword based approach to push blog posts to users. This is essentially a subset of our requirements because we also make use of social network information and user ratings to produce recommendations.

Some research that has focused on the scalability aspect of recommender system design includes [118, 119, 125]. However, our approach is different from the use of collaborative filtering as in Google news personalization [119], or the construction of artificial social networks based on user similarity as in Tribler [125], or the replication of information flow patterns on social networks as in [118]. These methods are only able to replicate average user behavior, whereas our approach uses sociological theory to understand the contextual reasons behind a users preferences, and can hence produce closer-fitted recommendations. Although many open problems remain to be solved, we are hopeful that our methodology of integrating elements of information science and sociology will improve the state-of-the-art in recommender systems for participatory media.

Chapter 9

Discussion and future work

Communication is the way we transmit and reproduce our lifeworld – Jurgen Habermas, The Theory of Communicative Action, 1987

We started Chapter 1 with the vision that computer-mediated communication has the potential to improve human development by providing low-cost, universal access to useful information in different parts of the globe. We then developed various mechanisms for such communication and described them in Chapters 2 through 8. This chapter concludes the thesis and outlines areas for future work.

In Section 9.1, we summarize the key contributions of the thesis. This is followed by Section 9.2 where we situate our work in the broad perspective of other contemporary research, startups, and non-profit organizations operating in the area of information and communication technologies for development; we identify similarities and dissimilarities in vision, and contrast our methods with alternative methods applied by these agencies. Next, in Section 9.3, we discuss a few ethical implications of our work, and of any work in the area of information search and recommendation in general. These ethical implications are centered around a fundamental insight that we came to realize: namely, that it is very hard to separate the political orientation or beliefs of system designers from the algorithms they design. Moving towards an increasingly automated algorithmic approach to information search and recommendation may be inevitable given the information growth rates of today, but it stands the risk of misguiding societies if systems are designed carelessly. We follow this with a preliminary suggestion of a few principles that system designers should keep in mind. Finally, in Section 9.4, we discuss avenues for future work leading from our research, demonstrating that the work initiated in this thesis provides for several additional detailed explorations and new ideas for interesting systems.

9.1 Contributions

In chapters 2 and 3, we talked about the nuts and bolts of transporting information to different parts of the globe. We showed that the delay tolerance characteristics of non-interactive applications can be used to design specialized systems that result in significantly lower communication costs. With OCMP, we highlighted the opportunistic communication possibilities enabled by multi-NIC mobile devices, and prototyped a Java-based middleware to optimize communication on these devices. We used the same principles in KioskNet to design delay-tolerant mechanical backhaul systems to provide Internet connectivity in remote rural areas. Both OCMP and KioskNet illustrate the need for a fundamental shift in thinking about networking protocols to make them suited for delay-tolerant connectivity. This includes the following:

- *Session-layer persistence*: Opportunistic connectivity by definition implies that transport layer disconnections may occur before an entire data transfer session can complete. Therefore, endpoints require session persistence to be able to resume downloads and uploads without having to recommence them from the start, as explained in Section 2.4.3.
- *Application support*: Support for delay tolerance requires new APIs to make it easier for software developers to program such applications. We proposed a few such APIs in Section 2.4.1.
- *Visibility of link-layer events*: In order to efficiently terminate transport layer sessions without waiting for connection timeouts, we showed in Section 2.6 that notification of link-layer events to higher layers on the same device, or over a control-channel from the correspondent device, can offer significant benefits.
- *Need for policy-driven scheduling*: When opportunistic connectivity is considered as an intermittently available resource for which different applications on a mobile device compete with each other, the need for scheduling algorithms emerges naturally. As explained in Section 2.4.8, a choice of schedules implies the need for policy-driven methods, for example, that users be allowed to define policies to optimize the monetary cost of communication, or power consumption, or delay in data delivery.
- *Mobility versus address aggregation*: Whereas address aggregation is possible if nodes taxonomically similar to each other are also topologically close, mobility implies that nodes with similar addresses need not be close to each other at all. This leads to an inevitable lookup step to translate from globally unique identifiers of nodes to their current location specific addresses. Section 3.4.1 explains this relationship, and Section 3.4.5 explains how lookups can be made scalable in a disconnected network.

- *Self-describing messages*: When multiple infrastructure elements share the responsibility for delivery of messages, scalability can be ensured if messages carry enough state so as to not impose additional work in consulting some specific infrastructure element. This insight is shared with IP networks where IP packets are made self-describing so that intermediate nodes not need consult the source or destination nodes to forward IP packets. This is explained in Section 3.2.

- *Security*: Traditional security architectures such as PKI are inefficient in the delay-tolerant scenario because they require a sender to fetch the public key of the recipient, which may involve a long round-trip to the Certifying Authority. As shown in Chapter 4, identity based cryptography provides a viable alternative.

Note that the communications systems of OCMP and KioskNet are relevant not only in the context of information recommender systems, but also to enable infrastructures for other systems that require connectivity. For example, KioskNet has found its use in the area of telemedicine where it has been deployed in Ghana to help doctors in rural areas exchange messages with highly trained doctors in urban areas.

We then directed our attention to the challenge of recommending useful information to users. In Chapter 5 we emphasized the benefits of sharing and creation of information in a participatory manner, and identified the components of context, completeness, and credibility to define the usefulness of information for different people. Chapters 6, 7, and 8 discussed algorithms to help participatory sharing and creation of information by recommending useful messages to users by building on the social networks within which users reside. We mathematically quantified the abstract concepts of context, completeness, and credibility, and used them in machine learnable models to learn user preferences. Our novel use of social networks to infer the social context of users showed that with availability of the right information, sociological insights can be adapted to develop useful algorithms.

The entire process we followed of (i) reasoning about the nature of useful information, (ii) the use of sociological insights to quantify the different components of usefulness, and (iii) the design of learnable models to infer user preferences towards these different components, illustrates the development of information recommendation algorithms from first principles. Some principles that formed the basis of this work are:

- *Multi-dimensionality of usefulness*: Information may be useful for different users for different reasons. Therefore, unless the construct of usefulness is atomized into its individual components, the reaction of users to information may appear to be random. In Chapter 5, we illustrated two components of usefulness, namely context and completeness, and a third component of credibility that is a property of contextual and complete information. Our research shows that identifying individual components is valuable if personalized information recommendation is to be achieved.

- *Contextual layout of social networks*: The structure of social networks seems to arise from certain underlying factors related to the ways in which humans form relationships. Sociologists have studied these factors over many years and contributed insights about how they manifest themselves into different social networks structures. We hypothesized in Chapter 6 that social context and homophily are important examples of such factors that are responsible for structuring social networks. We then validated our hypothesis by showing that social network graphs could be analyzed to infer and extract shared contexts of people. We also showed that information contributed by people sharing the same context helps them improve their understanding of various issues, while information contributed by people from diverse contexts helps provide a broader perspective to them.

- *Causal modeling*: Careful understanding of the relationships between various evidence variables and the constructs of context, completeness, and credibility, helped us form causal models such as the usefulness and credibility Bayesian networks in Chapter 7. We went over many iterations and careful correlation finding to identify the most appropriate models. This suggests that modeling the actual causal relationships between all the variables that influence users is a challenging process. However, causal models have the advantage that they assist in reducing the relationships that are studied in order to learn user patterns.

- *Multi-dimensionality of credibility*: Similar to the concept of usefulness, credibility is also a multi-dimensional construct. As shown in Chapter 7, breaking credibility into components of context and completeness helped us make more closer fitted recommendations than collaborative-filtering mechanisms that tend to reproduce average user behavior.

- *Scalability through distributed architectures*: The development of distributed architectures through careful overlay construction and partitioning of computing tasks presents a viable avenue to ensure the scalability of recommender systems. We verified in Chapter 8 observations such as topic-locality in social network graphs that can be used as an underlying principle to design topology construction and routing algorithms for scalable systems.

Our method of developing recommendation algorithms from first principles allowed us to incorporate many insights that we may not have considered had we tried to build on top of existing recommendation algorithms. We will add that our main focus is not to advocate the use of Bayesian networks or Eigenvector propagation or other specific tools that we may have used, but to emphasize the importance of operating from first principles that helped us incorporate sociological and other insights in developing algorithms.

9.2 Insights

In retrospect, we feel there were four guiding tenets that have formed the basis of our proposed systems. Although we unconsciously followed these tenets throughout, it was surprising to realize this, and even more surprising to realize their simplicity and obviousness.

- *Universal access* to information is critical to provide equitable growth opportunities to people.
- *Appropriate technology* is necessary to make the systems usable by the people. Here, we use the term "appropriate" with reference to writings such as *Small is Beautiful* by E.F. Schumacher [161], where it is advocated that technology is appropriate if it is designed keeping the needs of the people in mind, and their capabilities to use it, repair it, and innovate on it. In our case, low-cost is certainly one aspect of making the technologies appropriate. Further, the flexibility to access information via both cellphones and rural kiosks enhances the usability of our system to a wider diversity of users. However, there are more aspects that are necessary to ensure the appropriateness of technology, such as the social and organizational context of deploying and using the technology. Addressing many of these aspects is beyond the scope of this thesis.
- *Contextual information* is necessary to gain a better understanding of various issues.
- *Complete information* is necessary to form a broader and more informed perspective on various issues, and is becoming increasingly important as the world gets ever more interconnected through economical, environmental, and human ties.

Although discovering these tenets was both difficult and exciting, similar themes arise in other existing work. We next discuss a number of initiatives we came across that revolve around one or more of the same tenets described above. Not all of these initiatives focus exclusively on news related information, but they can be considered as building blocks or valuable lessons towards the vision we have stated.

- *Bridging the digital divide*: Many projects focus on the need to bridge the digital divide, to enable rural and marginalized communities to access the Internet and benefit from it. The One Laptop Per Child (OLPC) project is one such initiative that was conceived to develop a $100 laptop for children, with support for features such as ad-hoc WiFi networking for Internet connectivity, a hand cranked battery recharger, and non-reflective displays to avoid the glare of sunlight when used outdoors [72]. A recent redesign has lowered the cost to $75 by replacing the keyboard with a touch screen that can either function as a keyboard, or can be used to convert the laptop into an e-book reader [162]. The project brought in many technical innovations to keep the cost low, and to make the laptop robust and functional in the challenged environments of rural areas in developing countries. However, it is also surrounded by significant criticism, mostly emerging from questions about the appropriateness of the technology. Will the device provide any significant advantages over traditional methods of education that involve a slate and chalk, or even paper notebooks, supplemented by a traditional library? Do the teachers who work in these remote areas have enough technical expertise to enable efficient use of the technology? Will replacements and spare parts for the system be available readily?

Projects with a similar motivation but different strategy, and now gaining popularity in many countries, are those of telecenters or rural Internet kiosks [163]. As explained earlier in Chapter 3, these kiosks provide public access terminals from where users can access the Internet, or various services related to e-governance and e-commerce. The kiosk operators are trained to also help those customers who may not be sufficiently tech-savvy or educated enough to use the Internet themselves. Since these projects provide a shared infrastructure, they are certainly lower costing than the OLPC project. Involving the kiosk operator also puts a human in the loop which improves the usability of the kiosks. However, studies on the benefits of telecenters indicate mixed results [164]. Many telecenters face problems of Internet connectivity which impacts their usability. Further, it is hard to find well educated kiosk operators in rural areas who can be trained to offer specialized services – very often, the kiosk operators are capable to only help people access highly automated services such as bill payments, and not anything more intellectual such as find information related to agricultural advice.

Both the telecenter and OLPC projects focus on the tenet of universal access, but the appropriateness of the technologies is questionable when considered in the context of the people who are meant to use them. In our thesis, we have proposed methods to improve Internet connectivity that can be used for these projects, but the overall usefulness of the technologies also depends upon the supporting social and organizational structures necessary to execute the projects. We will face the same challenges when we try to deploy our proposed systems for information recommendation. We discuss later in Section 9.4.4 some initial ideas to move forward effectively on this goal.

- *World Wide Telecom Web (WWTW)*: This project is in its initial stages at the moment, and focuses on the development of voice driven interfaces to help illiterate and semi-literate populations to create and access information [165]. The project is primarily aimed to help small businesses run by individuals to gain an online presence to advertise about their products, list the services provided by them, and interact with customers. Tools are provided to build *voice-sites* analogous to Internet web-sites, and to browse across these voice-sites. Given the growing popularity of cellphones in the developing world, a voice-driven interface appears to be appropriate in such contexts. Ensuring usability of the interface is certainly challenging but can hopefully be addressed through relevant research and innovations.

This project can complement our system by allowing users to access information not only through data communication over cellphones and rural kiosks, but also through voice.

- *WiFi-based Long Distance Networks (WiLDNet)*: The WiLDNet project developed modifications to the 802.11 protocol to make it operate more efficiently over long-distance wireless links [166]. Deployments were done in India to provide telemedicine facilities in a number of rural clinics by connecting them to an eye-care center in a nearby city, and the staff at the clinics were trained to operate the system, cameras, and printers. While this was a good example of the application of technology to solve a social problem, the project team also conducted extensive analysis to ensure robustness of the equipment in the remote rural environments. For example, the grid power supply was found to be poor with extreme fluctuations, and harmed the equipment. Lightning strikes on tall towers, misconfigured routers, wrong wirings, etc were found to be other causes that resulted in system downtime [167]. The findings were supplemented with better circuit designs to ensure longer lifetimes for the equipment, and training manuals to make it easier for the local staff to repair the equipment.

  Although this project was deployed in the telemedicine context, the technology is certainly useful to provide telecommunication facilities in rural areas even for news related information systems. The learning derived from keeping the system design strongly grounded in the local context to make the technology appropriate, will certainly be applicable for our systems too.

- *Software based community radio systems*: Community radio involves operating a low-power AM/FM radio station covering a radius of $\sim$ 10-20km, where the local community of the region produces radio programs related to common issues of the region. Local production of content makes it highly contextual. This is especially useful in rural areas of developing countries where poor-education and lack of media agencies makes it hard to disseminate useful information to rural populations about health, entrepreneurship, education, etc. Furthermore, radio is a convenient medium to disseminate information because radio-receivers are cheap and owned by a lot of people.

  Gram-Vaani (*www.gramvaani.org*), an organization recently founded by us, aims to make the setup and running of community radio stations in rural areas cheaper, efficient, and more interactive. We plan to reduce the cost by eliminating hardware such as FM exciters and transmitters, and instead use software based systems for recording, mixing, and transmitting radio broadcasts. Software systems will also provide significant flexibility to enable novel interactive applications on the broadcast medium of radio. We will do this by making use of the Radio Data System (RDS) protocol that supports low-bandwidth transmission of digital information in parallel to the FM audio broadcast, and can be used to send control codes to radio receivers. The radio stations will also be connected to the Internet to allow radio stations to exchange content among each other, and also access other information provided by news agencies, agricultural research institutes, government agencies, and other entities. We plan to enable the information flow via the recommender system proposed by us in this thesis, where the personalization algorithms will not function at the level of each individual in a village, but at the radio station level to provide information to the radio station operators.

We have tried to incorporate all the four tenets in the vision behind Gram-Vaani. Low cost, voice based access, and the use of broadcast radio will make information dissemination possible universally, and in an appropriate manner for rural populations. Involvement of the local community in radio programming will make the information contextual. Internet connectivity to a recommender system will ensure completeness as well. This project will therefore serve as a first step in implementing some of the ideas proposed and prototyped in this thesis.

- *News and knowledge-sharing websites*: Participatory media and Web 2.0 websites have been rapidly gaining popularity. This includes websites such as Exit133, NewAssignment, TheHoot, OhMyNews, NewsVine, and Digg, each representative of a specific category of its own. All of these websites aim to address the need to provide either contextual information or complete information or both. We next discuss these websites to point out the wide diversity of mechanisms for knowledge sharing employed by them.

  Exit133 refers to itself as a *hyper-local* journalism website that caters specifically to local issues in the region of Tacoma in the state of Washington in USA. They welcome user contributions, pointers to articles published in newspapers, and discussions on various local topics. The locality of discussions helps make them contextual to the Tacoma population. Although the model is similar to that of local newspapers or the vernacular press, the interactivity provided through the online medium enables information evolution over time when more and more people share their thoughts.

  NewAssignment takes citizen journalism a step further by engaging professional journalists to investigate matters of concern to the local population, that are not investigated by the popular press. This combines the independence of citizen journalism with the thoroughness of professional journalism. Requests by people for “new assignments” helps understand the contexts in which people want the investigation to be carried out, a combination of multiple contexts improves completeness, and the professionalism of journalists improves credibility.

  TheHoot is a media watchdog in India. It scrutinizes the news published by various news agencies, and points out places of media bias or incorrect facts. Journalists, media professionals, and common people come together on the website to debate about various news articles. Such a centralized website to aggregate all comments and feedback improves completeness of the news under discussion.

OhMyNews follows a very interesting model of citizen journalism in S. Korea: people contribute articles, but an editorial board selects good quality articles, edits them, and gives monetary rewards to people whose articles are published on the website. This provides both completeness by welcoming views from a wide diversity of people, and credibility by having the editorial staff validate the submissions. OhMyNews is now extremely popular, and is often cited as being one of the most influential website to affect the results of the S. Korean presidential elections in December 2002.

Digg invites pointers to articles and blogs published by any source, but unlike OhMyNews where editors select articles for publication, Digg users vote to promote articles to the front page. Users are also allowed by comment on articles. The democratic voting procedure helps improve completeness in news about various topics, and comments bring about further information evolution.

NewsVine automatically aggregates articles about current affairs from a variety of news sources, and uses proprietary algorithms to select articles for publication based on their freshness, credibility, and popularity. This leads to a highly dynamic online news source that gets frequently updated with the latest breaking news. Users are also allowed to comment on the articles, resulting in greater completeness of views and opinions.

All these websites use different mechanisms but share a common goal to improve the provisioning of contextual and/or complete information related to news. The websites can be easily examined under the lens of context, completeness, and credibility defined by us in Chapter 5. The success of these initiatives lend validity to our goal of incorporating user participation to improve the effectiveness of news media. Although none of the websites currently support personalization, our methods can be easily adapted if the required information such as topic specific social networks and user ratings are available. In fact, most of these websites do support social networking and ratings, and we plan to test our personalization methods on more extensive datasets obtained from crawls of these websites.

- *Automated recommender systems for news*: As mentioned earlier in Chapter 7, Google News is one of the few examples of a personalized recommender system focused on news. Their approach to cluster together articles about the same story helps increase completeness by making it easier for users to discover similar articles with potentially divergent viewpoints. However, the degree of completeness is not quantified, and they do not make use of social networks to personalize information. Google News also does not record any ratings or comments by users, and relies exclusively on clicks to learn user preferences. This makes their methods very different from our own, even though they share a common tenet of helping to provide complete information.

Some of the works listed above focus on only a few tenets but not all. This impacts their validity and usefulness if they are examined in isolation. We leave it to the reader to think about how these initiatives complement each other, and can be clubbed together into a more holistic approach. We will also add that the four tenets of universal access, appropriate technology, contextual information, and complete information may not be exhaustive, but they certainly seem to be necessary for the design of any information system.

9.3 Ethical implications

The design of information systems raises fundamental questions about the prescriptive or descriptive objectives of the system designer. A purely descriptive objective, as what we have tried to adopt in this thesis, would recommend information to people that they are highly probable to like. However, a more prescriptive objective, if the system designer takes on the role of a teacher, would involve purposive action on the designer's part to indoctrinate certain beliefs into the users. This is possible because the system designer commands a highly influential position by holding keys to algorithms that can control which information is filtered out or pushed to the users, as with popular search engines like Google and Yahoo. Even having a descriptive objective is not an entirely safe approach because it stands the risk of reinforcing existing beliefs of people without encouraging them to expand their horizons.

The drawbacks of both descriptive and prescriptive objectives seems even more severe when the emergent behavior of information systems is considered. For example, information recommendation may actually cause users to form new social network ties or drop existing ones, definitely a critical step because it changes the contexts in which people operate in the real world. It therefore seems that it is impossible to separate the political orientation of the system designers from the algorithms they design. What is even more serious is pointed out by this quotation:

Machiavelli undermines classical politics by changing opinion as a representation of belief, conviction, or conduct into an appearance strategically sculpted to make a seamless convincing impression – Thomas Goodnight on Habermas

Algorithms can be crafted to actually prevent rational critical debates as advised by Habermas, and instead strategically manipulate users into believing certain impressions. As computation capabilities grow, and research into sociological, psychological, and biological processes of humans reveals more and more insights into cognition, this becomes even easier to implement. Thus, one of the greatest risks that societies may face as information access takes a more automated and algorithmic approach, could arise if system designers adopt a Machiavellian perspective towards furthering their own beliefs. This fear is similar to the classic Orwellian scare on the possibilities of mind control exercised through news managed by autocratic governments [168].

One way to avoid this is by letting people manage the algorithms themselves, based on the democratic principles of *transparency* and *feedback*. This can possibly be done by mandating a monitoring framework that every information search and recommender system must provide. This means to design intuitively understandable metrics such as context and completeness based on insights revealed by sociology and media research, that can help explain the behavior of the information system. These metrics should then be transparently exposed to the people, and feedback should be gathered to modify the system if so desired by the people. A monitoring infrastructure that supports transparency and feedback is critical to any democracy, and we feel the same principles should also be employed before algorithms are allowed to be unleashed on users.

9.4 Future work

While we have provided several valuable contributions and insights for the problem of providing universal, low-cost access to useful information for users across the globe, there are also several problems that merit further exploration in the future, as described below.

9.4.1 Opportunistic communication

Cross layer interactions during opportunistic communication

The rapid establishment and teardown of connections during opportunistic communication leads to interesting cross-layer interactions that have a significant impact on performance. [65, 169] identified various stages during the lifetime of a drive-thru WiFi connection. The first stage is the connection establishment stage that includes a WiFi association step followed by DHCP address acquisition. Both these steps involve a number of packet exchanges between the client and access point. It was found that even one dropped packet during this stage can some times trigger a complete recommencement of the entire connection establishment process. This can lead to wastage of valuable connection opportunities, and there seems to be room to innovate for more optimized protocols.

The second stage is the actual data transfer stage. This involves the negotiation of an optimal transfer rate at the MAC layer, and efficient utilization of the available MAC rate at the TCP or alternate transport layer. Rate negotiation can be quite complicated especially during a drive-thru scenario, because the RSSI values change rapidly with varying distances of a moving wireless client from a fixed road-side accesspoint. [169] suggested to use fuzzy-logic to learn the profile of a wireless link at different distances from the access point, and utilize the model to change the rate for a moving client. Even if optimal rate negotiation does happen, the ability of the transport layer to utilize it is also challenging. This is because the transport layer operates on an end-to-end basis, and therefore the flow- and congestion-control algorithms will be affected not just by the last hop wireless link, but also by other network characteristics along the route. Although this problem can be mitigated by split-TCP or buffered access points, an additional issue is the different reaction times of the transport and MAC layers to adjust their rates. This probably calls for fine-grained cross-layer information exchange between the MAC and transport layers.

The third stage is the connection breakdown stage, and again requires careful timing to disconnect at the right time so as to not waste connection opportunity for the client, while not wasting the resources of the accesspoint either. We found that delayed breakdown of the transport layer connections could lead to extra state maintenance at both the endpoints, and could be improved by message exchanges on an out-of-band control channel. This shows that closer investigation of opportunistic connection sessions is required to improve their utilization.

Mobility prediction and scheduling

The need for optimal scheduling algorithms was outlined in Chapter 2, to allocate wireless resources in a fair manner to different users and applications, while preserving constraints such as delivery deadlines, low communication costs, and low power usage. Assuming perfect future knowledge of connectivity intervals, [170] proposed a fast utility-based scheduling algorithm that was provably optimal. However, perfect future knowledge is practically impossible to possess. Even though advances have been made in better mobility prediction algorithms by making use of calendaring information [171], social-networking features [37], etc, mistakes in prediction are inevitable. We therefore feel that there is a need to design scheduling algorithms that are not just approximately optimal, but also robust. We are not aware of any related work that handles the problem of robustness in scheduling, but we believe that research in statistical decision making will be relevant in this regard [172]. The Dempster-Shafer theory of reasoning seems more suitable in the context of robust scheduling than Bayesian probabilities because of its ability to incorporate confidence measures in belief functions.

Use of mobile devices as sensors

With the massive proliferation of mobile devices, in the past we have toyed with the idea of using them as sensors to monitor wireless deployments [186]. Mobile devices can passively record data such as WiFi signal strengths, 3G bandwidths, etc, at different times of the day and different GPS-reported locations. The data can be transmitted to a central establishment where it can be collated on spatio-temporal databases to infer maps of cellular and WiFi data transfer performance. Now, if a mobile device is looking for wireless networks, or trying to decide the best network to use, it can simply query this database on the control channel and make more informed decisions for policy control and scheduling. The growing presence of mobile devices and their increased outreach due to mobility, indeed makes this a practical avenue to pursue. Issues such as privacy do arise, but cryptographic protocols can be designed to anonymize the data that is recorded. There is a similarity in vision of this idea with other projects on participatory sensing using cellphones, where cellphones are fitted with pollution sensors to develop pollution maps of cities [173]. There may be room for exchange of research findings, especially in the area of spatio-temporal databases.

9.4.2 Social network theory

Understanding social network patterns

As we showed in Chapter 6, social context appears to be a significant underlying factor that contributes to structuring social networks. However, we have not as yet investigated in detail the network structures that arise in different scenarios, for example, ways in which the network structure of college communities may be different from that of information seeking communities. Characteristics of the distribution of strong and weak ties in various topics (Section 6.3.2), may give hints on how communities related to these topics evolve, and the ideal connection patterns that help communities to grow [174, 175]. Finally, such findings can be used to develop mathematical models that explain the observations. For example, [96] proposed a model to explain the navigability and small-world property of social networks, [176] proposed a model to explain the shrinking-diameter property of social networks, and [177] proposed a model to explain the scale-free nature of social networks. In a similar way, models can be developed to explain the connectivity patterns for different topics, and additional observations such as those made in Chapter 8 related to topic locality and popularity. Such a model will make it easier to discover further interesting properties that may occur, and also help in designing better algorithms that make use of these properties.

9.4.3 Social network applications

Recommendation and ranking algorithms

We mentioned in Chapter 7 many limitations of the recommendation algorithms proposed by us, such as the problems of cold-start, attack resilience, and the modeling of confidence bounds. We also made a number of assumptions about the validity of the datasets used for the evaluation, the choice of clustering algorithm, suitability of the metrics defined by us, and the requirement for the collaborative algorithms to have knowledge of message ratings from a large and common set of users. A number of avenues for future work were also mentioned, such as by enhancement of the credibility model to include expert- and role-based credibilities, and experimentation with a hybrid algorithm to combine the usefulness and credibility algorithms. In the future, we plan to investigate these details and to conduct evaluation tests on more extensive datasets. A problem we definitely foresee is to find suitable datasets for our experiments, because most websites such as Digg, Newsvine, etc, do not explicitly differentiate between credibility and usefulness ratings. Therefore, there might be merit in collecting datasets by actually developing our own pilot recommender system with a few users. Tools are now becoming available to do such pilot tests without much difficulty. For example, APIs are now available through Facebook and the OpenSocial initiative to design applications and get information about social networks of users. The flexibility provided by browsers including Firefox and Flock to integrate with client-side applications, has made it easier to solicit user feedback or observe user reactions such as the time spent by users on a particular message. Finally, the recently released semantic-web-based OpenCalais service by Reuters to generate RDF metadata of blogs and news articles, makes it convenient to use topic categorization services for rapid innovation. We will use these tools to develop and test a pilot system based on our algorithms.

Exploration-exploitation tradeoff

The hybrid algorithm and associated routing infrastructures outlined in Chapter 8 aim to replicate the observed user behavior. However, the system should also be capable of helping the user to explore new territories – topics, opinions, etc – if the user so desires. This is referred to as the *exploration-exploitation tradeoff* [145], also mentioned in Table 5.2. We feel that the reflector component our proposed architecture can be used to enable this feature. Exploration of unfamiliar clusters can be done by adding temporary routes across reflector nodes not expected to be similar to each other. The routes can then either be reinforced based on ratings given by users in either cluster, or converted into real routes if the users begin to explicitly declare ties to each other. Virtual routes to be added can in fact possibly be inferred based on link prediction and collaborator search algorithms [160]. Messages propagating across these virtual routes can then either be automatically pushed to users, or an interface can be provided where users are given an option to view messages from clusters they have not seen so far. We will explore this avenue in future work.

Data structures and API design for social networks

The design of efficient data structures for social networks is a very new research area. The only work we have come across is [150] to design data structures that make it convenient to compute metrics such as edge betweenness and the centrality of nodes. However, there was no taxonomical basis that was discussed as to whether some particular methods are more suitable for certain types of tasks instead of other methods. Forming a more organized perspective will be essential to design generic data structures that can be known to work efficiently for certain types of tasks.

This brings us to the question of defining various tasks for which data structures should be optimized. We outlined many tasks in our recommendation algorithms such as clustering of the topic specific social network graphs, computation of clustering and integration coefficients, Eigenvector calculations, etc, that need to be performed in the data center. It may be hard to optimize data structures for such high level operations, and therefore certain primitive API calls need to be identified that can be composed together to compute complicated metrics. We are not aware of any work so far on APIs for querying social network graphs. Note that although simple APIs like querying for a list of friends are provided by Facebook and the OpenSocial initiative, we are more interested in APIs that make it easier for third party developers to compute complex metrics or run algorithms on the social network graph. This is because the network graph is already owned and hosted by these organizations, and therefore they are in a better position to also host the required computation infrastructure to allow third party developers to write more algorithmic and programmable applications such as the information recommendation application we have proposed.

9.4.4 Appropriate user interfaces for developing regions

We assumed in the use-case discussed in Chapter 1 that the farmer was literate and owned a mobile device with which he could access the recommender system. This is however a very strong assumption because in most developing countries, such farmers may not be sufficiently educated to use modern computerized devices, or even be rich enough to afford one. Therefore, appropriate user interfaces need to be developed to include more people and make the system usable by them.

We use the term *user interface* here in a very generic sense to indicate the medium through which a human being may access information; this could include novel projects such as reading messages published automatically as a personalized or community newspaper [1], or listening to messages over a radio broadcast [2], or

[1] Print casting, California, *http://printcasting.com/*: Blogs and articles from personalized RSS feeds are automatically typesetted into a newspaper to also involve those people into online citizen journalism who are mostly comfortable with the print medium.

[2] Community radio, India, *http://gramvaani.org/*: Radio jumps literacy hurdles in providing information even to poorly educated people, while community participation also ensures that the broadcast content is contextual.

accessing them through a call center [3], or reading them over the Internet in a kiosk or on a mobile phone [4], etc. Of course, certain media are more suited for certain social and cultural settings, or for viewing certain kinds of messages, or for solicitation of user feedback such as comments and message ratings, or for the management of social networks, etc, but the nature of the medium itself is superfluous to the actual flow of information.

Just like the Internet connected computers together, there is now a need to bridge the flow of information across different media so that people can be provided information access over the user interface they find to be the most appropriate to them. It is imperative to realize that only providing Internet connectivity is not the solution to bridging the digital divide, but providing connectivity through appropriate user interfaces is the essential requirement. This bridge to enable information flow across different media can either be through automatic transcoding of content across different media, or through manual contextualization as information is made to flow across; our belief is that the mechanisms are immaterial as long as the networks of information flow and evolution are considered, because information itself is independent of the medium that carries it, and only depends on the people that create and modify it.

[3] Freedom Fone, Zimbabwe, *http://www.kubatana.net/*: In a country where the print, radio, and television media are severely controlled, mobile phones are a suitable medium to spread information that inspires people and empowers communities.

[4] Voxiva, Peru, *http://www.voxiva.com/*: Far flung villages across high mountains may be inaccessible through conventional transportation, but electronic connectivity has helped bring these villages into the mainstream of a globalized and connected world.

Appendix A

Background and glossary

WiFi hotspots

WiFi, or 802.11, is a short-distance wireless technology operating in the unlicensed 2.4GHz and 5GHz bands. Many coffee-shops, restaurants, libraries, office and educational campuses, etc provide Internet access through WiFi. These regions of connectivity are commonly referred to as “hotspots”.

GPRS, EDGE, UMTS, 1xRTT, EvDO

These are standards for wide-area wireless data technologies for cellular networks. They are normally classified as 2G, 2.5G, or 3G technologies. GPRS is a 2G technology, EDGE and 1xRTT are 2.5G technologies, while UMTS and EvDO are 3G technologies.

Bluetooth

Bluetooth is a very-short-distance wireless technology meant for device-to-device communication such as between a cellphone and its headset, or between a keyboard and a computer.

PDSN and GGSN

These are gateway nodes in a CDMA and GPRS network respectively, that interface between IP connectivity on one end, and the standards specific radio-access-network (RAN) on the other end. PDSNs and GGSNs are the final infrastructure elements in a cellular data network that support IP packets. Transcoding services by companies such as ByteMobile often operate at the PDSN/GGSN level, or at some higher IP aggregate point. Therefore, these elements are ideal places to host OCMP proxies.

RTT: Round Trip Time

The long-term TCP throughput between two Internet endpoints is inversely- proportional to the round-trip-time for IP packets between the two endpoints [178]. Typical RTT values between the North American east and west coasts $\sim$ 50ms, but can often be as large as 500ms for inter-continental routes. Therefore, it is useful to maintain globally distributed caches or OCMP proxies to ensure low RTT values to mobile devices.

IMSI: International Mobile Subscriber Identity

Each cellphone has a unique identity that is initialized by the mobile network operator in the SIM card of the device. The IMSI contains information about the country, mobile operator, and a unique identifier.

SIM cards

A SIM (Subscriber Identity Module) card is a removable smart-card placed in cellphones and stores a variety of information such as the IMSI, phone addressbooks, and text messages.

DTN: Delay Tolerant Networking

DTN is a style of networking that enables store-and-forward data transfer even if an end-to-end connection between two endpoints is not available. Such scenarios typically arise in challenged environments like inter-planetary communication, sensor networks, underwater networks, and connecting remote rural areas to the Internet. The DTN Research Group (DTNRG) is engaged in researching architectural and protocol implications of DTNs, and to enable interoperable communications between such heterogeneous networks. Some concepts developed in [35] that we have used in our system design are the following:

- *Regions*: Regions are defined as a collection of DTN nodes within the same administrative boundary, or following the same communication protocols, or the same naming conventions, etc. For example, a collection of wireless sensors in a field may be referred to as a region because these sensors would communicate with each other on the same wireless protocol. Or, robotic probes operating on Mars may be grouped under the same region.
- *Gateways*: Gateways are DTN nodes that interface between two or more regions. For example, a typical gateway node in a sensor field aggregates readings from various sensors, and pushes the results to the Internet. Similarly, probes on Mars may send their data to a powerful gateway transmitter, that would connect to the Earth or some satellite to relay the data to space agencies.

- *Custodians*: These are DTN nodes that store data in transit if a connection to the final destination is not available. For example, if the gateway node in a sensor network does not have permanent Internet connectivity, then it would retain the data and transmit it only a connection becomes available. In this case, the gateway can also be referred to as a custodian for the data.
- *DTN links*: Many different types of links between DTN nodes have been categorized. *Permanent links* are those links that are always available, such as fixed wireless links between stationary nodes. *On demand links* can be established as and when desired, such as dial-up connections. Links may also be available at only certain *scheduled* times, or may be created *randomly*. Another category is that of *opportunistic links*, such as ad-hoc WiFi links that are created when moving nodes come in close proximity of each other.

As we show in Chapters 2, 3, and 4, delay tolerant networks introduce a different set of problems in connection management, routing, security, etc, that require a rethink of the conventional end-to-end networking model.

Rural kiosks

Rural kiosks in developing countries provide a variety of services such as birth and death certificates, bill collection, email, land records, and consulting on medical and agricultural problems. They are well-suited to this purpose because both the capital and operational expenses of the kiosk are spread among a fairly large user base, greatly reducing the per-user cost. Even with very low user fees (10-15 cents/transaction), an entrepreneur can make enough money to profitably provide government-to-citizen and financial services. The Government of India is supporting an initiative called *Mission 2007* to set up 100,000 kiosks in India through partnerships with private companies.

Orkut

www.orkut.com is a social networking website owned by Google. Similar to Facebook, users on Orkut can create a personal profile describing their interests, hobbies, profession, etc. Users can also declare explicit *friendship* links to other users, and are advised to accept links from only those users whom they know in real life. This is different from the notion of a link between MMORPG (Massively Multiplayer Online Role-Playing Games) gamers such those on Second Life, where a link need not necessarily represent a prior real-life relationship. In fact, Facebook makes it mandatory for users to confirm each others identity as their affiliations with schools, colleges, companies, etc. Orkut allows more flexibility, but our personal experience has shown that the Orkut social network is indeed a reflection of the real life social network. In general, the etiquette of most social networking websites including LinkedIn, Bebo, and Ryze promote identity disclosure and real life relationships.

We crawled the Orkut social network graph of more than 42,000 users for our experiments. Our reason for choosing Orkut was that the social network of all users is publicly visible to other Orkut users, and can hence be crawled for analysis. Orkut also allows users to form communities of interest and hold discussions on various topics. We crawled these discussion forums as well, which gave us information about the discussion content and the underlying social network of the participating authors and readers.

Digg

www.digg.com is a knowledge sharing website. Users submit interesting articles or blogs available on the Internet, and other users vote for these *stories.* Stories which gain a substantial number of votes are promoted to the front page of Digg. Users can also form a social network on Digg, much like the Orkut social network. However, the criteria on Digg for declaring an explicit link to another user may not be the existence of a real-life relationship to the user, but just an interest in reading stories by the user. Even so, the social network reveal information that we need about the context of users.

We used a crawl of the Digg website available from [134], which gave us information about the ratings given by users to stories, story submissions, and the social network of the users.

Friends, ties, links, edges

During discussions about our experiments, we use the term *friend* to denote an explicitly declared friend on Orkut or Digg, as the case may be. We synonymously use the terms *tie*, *link*, or *edge* to denote the existence of a relationship between a pair of users.

Graph clustering

The social network information can be represented as a graph, where users are nodes, and links between users are edges in the graph. We then use *graph clustering* algorithms to find clusters in the social network graph. A cluster is defined as a connected subgraph that has a high density of edges internal to it, and a relatively sparser existence of edges outside it. Graph clustering algorithms are able to find such clusters in the social network graph, which we subsequently use in our analysis for tie classification.

Content analysis

This is a term used heavily in media research and refers to manual or automatic analysis of content of newspaper articles, video shows, advertisements, political speeches, etc, which are referred to as *messages*. A typical content analysis task involves the labeling of messages based on some predefined taxonomy.

Recommender *or* recommendation system

A recommender system is an automated system designed to recommend relevant information to users. A good example of a recommender system is Amazon. *Users who bought this book, also bought...* is an example of a recommendation provided to a user. Amazon also keeps track of books purchased by users in the past, and uses this history to learn the preferences of users and accordingly make recommendations. Recommender systems have existed for movies and music, and are now even being considered for news recommendation.

Content-based recommendation

Recommendation algorithms generally come in two flavours: content-based and collaborative-based algorithms. Content-based algorithms extract features from the content of messages, and learn user behavior in terms of these features. When a recommendation decision has to be made about a new piece of content, the same features are extracted from the content, and evaluated on the learned user behavior model to predict the reaction of the user upon receiving the message.

Collaborative-based recommendation

To make a recommendation decision for a user, collaborative-based recommendation uses ratings for messages given by other users. Most algorithms operate by finding users similar to each other in terms of the message ratings given by them to old messages. Once similar users have been identified, ratings given to a new message by these users are then used to produce a final recommendation decision for the user under consideration.

Bayesian network

A Bayesian network is represented a directed dependency graph over some set of random variables, where a dependency edge between two variable indicates that the target variable is dependent upon the source variable. Bayesian networks allow the specification of models for causal relationships between variables, and have become a popular machine learning tool in the last few years.

Learning and inference

Machine learning typically involves two steps: A learning step, where patterns are discovered in the available data and relationships between these patterns is understood; and an inference step, where certain patterns are given, and the learning is used to predict the rest of the patterns. A common learning algorithm for Bayesian networks is the EM (Expectation-Maximization) algorithm, and common inference algorithms include MCMC (Markov Chain Monte Carlo) and Join Tree methods.

Datasets for evaluation, training *and* test datasets

Evaluation of machine learning algorithms are performed on datasets. Datasets are partitioned into *training* and *test* datasets. Training datasets are used in the learning step to discover relationships between various variables in the dataset. Test datasets use the learned model to make predictions, and compare the predictions with the actual values present in the dataset. If the predictions agree to a large extent with the actual values, it indicates that the learning algorithm performs well.

Jacquard's coefficient

The Jacquard's coefficient is used to evaluate the degree to which two sets A and B are similar to each other, and is calculated as $\frac{|A \bigcap B|}{|A \bigcup B|}$. This gives a value between 0 and 1. In our case, we use it to evaluate the degree of similarity between users over the sets of messages rated by them.

Principal Component Analysis (PCA)

PCA is often used as a dimensionality-reduction technique to reduce large datasets to its principal vectors that retain those characteristics of the datasets that contribute most to its variance. Lower dimensional analysis of these vectors can be done much easily than analysis of the original datasets. In Chapter 8, we plan to use PCA to calculate identity vectors for each reflector that retain the similarity or dissimilarity characteristics between reflectors.

Appendix B

Content-analysis based hypothesis validation

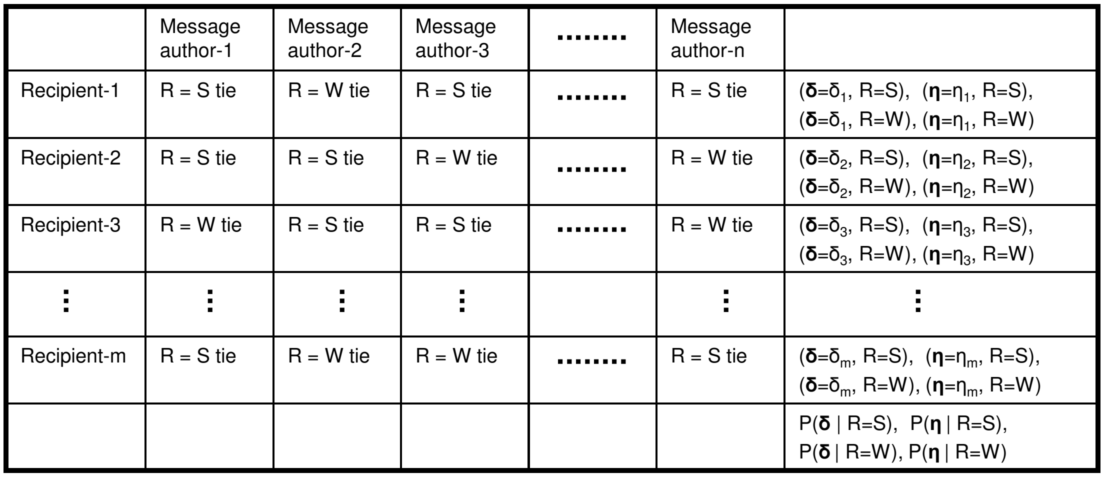
The hypothesis validation described in Chapter 5 can be enhanced by considering an alternative method of using content analysis to estimate η and δ as the context- and completeness-providing characteristics respectively of ties of recipients. In this appendix, we outline such a method.

To test the context and completeness hypotheses proposed in Chapter 6, a cohort of message authors and recipients needs to be first identified, so that given the nature of tie between message author and recipient (independent variable), the context and completeness provided by messages for the recipient (dependent variable) can be found. Fig. B.1 shows the design framework that can be used for the study [5]. The following random variables are defined:

- $\mathbf{R} = \{\textit{strong}, \textit{weak}, \textit{undefined}\}$, denotes the nature of tie between a message author and recipient.

- $\boldsymbol{\eta} \in (0, 1)$, denotes the context provided by messages to a message recipient. It is used to infer the distribution of another variable, $(\boldsymbol{\eta} \mid \mathbf{R})$, as the context provided by messages to a message recipient, given the nature of tie between message authors and the message recipient. Referring to the variables defined in the previous chapter, an instance of $\boldsymbol{\eta}$ is the average context $E[\eta]$ provided to a recipient r. For ease of exposition, we use η_i to denote the average context provided to the i^{th} recipient.

[5] Note that although this study falls under the heading of *message effects research*, it is different from most previous studies [13]. We use the nature of tie between a message author and message recipient to define categories (independent variable) of strong and weak ties, which produce two kinds of effects (dependent variables) on the message recipients, namely, to provide context and completeness. Previous studies have defined categories based on the type of message, for example, effects produced by textual content versus audio content versus video content. That is, the categories were not dependent on the tie between message authors and recipients.



	Message author-1	Message author-2	Message author-3	....	Message author-n	
Recipient-1	R = S tie	R = W tie	R = S tie	....	R = S tie	($\boldsymbol{\delta}$=δ_1, R=S), ($\boldsymbol{\eta}$=η_1, R=S), ($\boldsymbol{\delta}$=δ_1, R=W), ($\boldsymbol{\eta}$=η_1, R=W)
Recipient-2	R = S tie	R = S tie	R = W tie	....	R = W tie	($\boldsymbol{\delta}$=δ_2, R=S), ($\boldsymbol{\eta}$=η_2, R=S), ($\boldsymbol{\delta}$=δ_2, R=W), ($\boldsymbol{\eta}$=η_2, R=W)
Recipient-3	R = W tie	R = S tie	R = S tie	....	R = W tie	($\boldsymbol{\delta}$=δ_3, R=S), ($\boldsymbol{\eta}$=η_3, R=S), ($\boldsymbol{\delta}$=δ_3, R=W), ($\boldsymbol{\eta}$=η_3, R=W)
⋮	⋮	⋮	⋮		⋮	⋮
Recipient-m	R = S tie	R = W tie	R = W tie	....	R = S tie	($\boldsymbol{\delta}$=δ_m, R=S), ($\boldsymbol{\eta}$=η_m, R=S), ($\boldsymbol{\delta}$=δ_m, R=W), ($\boldsymbol{\eta}$=η_m, R=W)
						P($\boldsymbol{\delta}$ | R=S), P($\boldsymbol{\eta}$ | R=S), P($\boldsymbol{\delta}$ | R=W), P($\boldsymbol{\eta}$ | R=W)

Figure B.1: Design framework for study

- $\boldsymbol{\delta} \in (0, 1)$, denotes the completeness provided by messages to a message recipient. It is used to infer the distribution of another variable, $(\boldsymbol{\delta} \mid \mathbf{R})$, as the completeness provided by messages to a message recipient, given the nature of tie between message authors and the message recipient. Referring to the variables defined in the previous chapter, an instance of $\boldsymbol{\delta}$ is the average completeness $E[\delta]$ provided to a recipient r. For ease of exposition, we use δ_i to denote the average completeness provided to the i^{th} recipient.

Knowledge of the distribution of $(\eta_i, R = \{S, W\})$ and $(\delta_i, R = \{S, W\})$ for all recipients will eventually allow us to infer the probabilities P$(\boldsymbol{\eta} \mid \mathbf{R})$ and P$(\boldsymbol{\delta} \mid \mathbf{R})$. This will indicate the context and completeness provided by the two different types of ties over the entire population of message recipients. The completeness hypothesis will be validated if P$(\boldsymbol{\delta} \mid \mathbf{R} = \mathbf{W})$ has more of its mass to the right than P$(\boldsymbol{\delta} \mid \mathbf{R} = \mathbf{S})$. This can be evaluated using standard z-tests or t-tests to compare the mean of two probability distributions, and will show whether or not strong ties of a recipient provide less completeness than parts of the social network connected through weak ties to the recipient. Similarly, the context hypothesis will be validated if $P(\eta|R = S)$ has more of its mass skewed to the right than $P(\eta|R = W)$. This will show that strong relationships of a recipient provide more context than the recipient's weakly connected parts of the social network.

Note that such a framework is defined for each broad topic that will be examined. The labeling of relationships as strong or weak will be done by algorithms for identification of clusters in each topic specific social network. Quantification of η_i and δ_i will be done based on the experimental methodologies described later. The steps to test the hypotheses are as follows:

1. Assemble a cohort of message authors and readers willing to participate in

the study. Extract the topic specific social networks of the cohort for a few topics.

2. Identify clusters within these topic specific social networks.

3. Test the validity of the model using content analysis for the completeness hypothesis and surveys for the context hypothesis.

Cohort identification

The traditional approach to identify a cohort in a social network is to use snowball sampling: start from a single person, and then use name generators to expand the network to friends of this person, friends of friends, and so on [179]. A list can then be assembled of all topics these blogs cover. For example, climate change, or the Iraq war, or poverty eradication, etc. A few of the most popular topics can be selected, and for each topic, blog authors and recipients in the cohort interested in that topic can be identified. Algorithms for identification of clusters can now be used to label the ties between the authors and recipients as strong or weak. Eventually, matrices such as the one shown in Fig. B.1 can be assembled for each topic.

Clearly, using manual name generators for a large cohort can be very time consuming. An alternative is to extract the social network by crawling social networking websites that the participants would be using. A limitation of automatic social network extraction over manual name generators however can be that of missing data and spam data, if the online social network of a participant does not reflect the actual physical social network. [180] however found that this was not a significant problem for most users, and that online and real social networks of people indeed coincided to a large extent.

Identification of clusters

Once the topic specific social networks have been extracted, different clustering algorithms can be used to cluster the graphs [104–106] and label relationships as strong or weak or undefined.

Completeness hypothesis

A single message is taken to be the unit for content analysis. The goal is to determine $(\delta_i, R = S)$ and $(\delta_i, R = W)$ for the i^{th} recipient, as the average completeness provided by the set of messages written by authors linked through strong or weak ties to the recipient. 5 messages from each author will be considered.

Content analysis in the social sciences has been traditionally done by developing a set of rules through which coders can analyze and label the characteristics of content they examine [181]. The reliability of content analysis is determined by calculating the Kappa coefficient for inter-coder agreeability. Low agreeability on pilot tests suggests that the rules were not interpreted consistently by the coders, and there is a need to state the rules more precisely.

Precise rules can be stated through ontologies [182]. An ontology for a topic expresses the relationships between various aspects relevant to the topic. A sample ontology for cotton farming in India was shown in Fig. 5.3. Once ontologies are developed, rules for content analysis can be stated in a straightforward manner in terms of each node of the ontology. For example, do the messages discuss the role of free-trade in determining cotton prices, do the messages discuss the role that commodity exchanges can play in smoothing the global cotton price fluctuations, etc. The determination of $(\delta_i, R = \{S, W\})$ now becomes simple. For a set of messages, δ_i can be stated as the fraction of the area of the ontology covered by the messages, giving equal importance to each node of the ontology. Referring to the a_e and a_M variables defined Chapter 6, $\delta_i = \dfrac{|a_M|}{|a_e|.|M|}$. Therefore, the completeness hypothesis effectively implies that the fraction of area covered by messages written by authors in the same cluster, tends to be less than the fraction of area covered by messages written by authors in different clusters. In other words, the hypothesis states that strong ties tend to focus on the same matters repeatedly, but weakly connected parts of the social network provide non-redundant information and diverse views that touch upon other related aspects as well. Knowledge of $(\delta_i, R = \{S, W\})$ for different recipients will generate the distribution for $(\boldsymbol{\delta} \mid \mathbf{R})$ to test the hypothesis.

Context hypothesis

The context hypothesis is hard to test through content analysis alone. The value of $(\eta_i, R = \{S, W\})$ for the i^{th} recipient will depend upon the context of the recipient; therefore, it would be invalid for an external observer to instrument this value without being in the same context as the recipient. Knowledge based surveys to test for the understanding gained from different messages are also hard to do because of the threat of maturity: recipients would tend to read messages more closely if they are aware that they will have to later fill out a survey [183].

A suitable technique is as follows. Ask the participants to rate messages on a 5-point Likert scale (1=low, 5=high) based on their self-assessment of context promoted by the messages, but ensure that all of them use the same criteria in their assessments. This criteria can be based on a few examples and thumbrules that should be informally discussed beforehand with them. For instance, the following thumbrules can be considered for Example-1 given in Chapter 1 about the Emergency in Pakistan:

1. Does the blog entry refer to how the Emergency may impact your lifestyle? For example, its effect on your job, or your safety in going to work, or the prices of groceries? If the blog entry talks about such issues that would be relevant for you and your family, then you may want to rate this entry higher. However, if the blog entry talks about issues unrelated to you, then you may want to give it a lower rating.

2. Did you understand the main points that the blog entry was trying to convey? If so, then you may want to give it a higher rating. However, if the blog entry was completely incoherent, then you may want to give it a lower rating.

3. Did the blog entry sufficiently simplify the arguments it was trying to make? For example, if the author cited articles from economics or political science research journals that discuss issues relevant to the Emergency, then did the author simplify the conclusions of these research articles and their relevance to the event? If so, then you may want to give this entry a higher rating.

Suppose now that a Likert scale rating of j was given by the i^{th} participant to s_{ij} messages by strong relationships and w_{ij} messages from weakly connected parts of the social network. $(\eta_i, L = S)$ can now be estimated as follows:

$$u_i = \frac{\sum_{j=2}^{5} s_{ij}(\sum_{k=1}^{j-1} s_{ik} + w_{ik})}{\binom{m}{2}}$$

Here, $\binom{m}{2}$ is the total number of pairs of messages. Thus, η_i is the fraction of the number of pairs of messages by strong ties that promoted more context than other messages. $(\eta_i, R = W)$ can be calculated similarly, by reversing the values of s and w in the equation above. Knowledge of $(\eta_i, R = \{S, W\})$ for different recipients will generate the distribution for $(\boldsymbol{\eta} \mid \mathbf{R})$ to test the hypothesis. Note that this formulation of η_i does not reflect the actual context provided by messages, but for testing the context hypothesis, there is clearly a one-to-one correspondence between the validity of this rank based formulation of η_i, and a desired formulation which could measure the actual context provided by the messages.

This method does not suffer from the threats of validity and reliability. The same set of thumb-rules will be given to all participants for rating messages based on the amount of context promoted by them. This will provide reliability to the results because it will ensure that all the participants will use the same consistent definition of context which needs to be measured for this test. Validity of the tests will also be ensured because we will only consider the relative amounts of context promoted for different recipients by the messages. Therefore, the instrumentation errors that may occur because of inherent differences among the participants, such as their history or maturity, will not affect the results.

We will pursue this method in the future. This method was not a first choice for us because it requires significant time and resources for a thorough investigation.

However, following the success of the “lighter” method used by us so far, we now feel confident to undertake this study.

Appendix C

Sample surveys

C.1 Tie classification

Sample survey

Please rank your following 5 friends on a scale of 1-5 (1=acquaintance, 5=very good friend) in terms of how close they are to you and your immediate circle of friends.

1. vijay: *http://www.orkut.com/Profile.aspx?uid=...*
2. ABANI: *http://www.orkut.com/Profile.aspx?uid=...*
3. Tushar: *http://www.orkut.com/Profile.aspx?uid=...*
4. Seshadri Kiran: *http://www.orkut.com/Profile.aspx?uid=...*
5. Prabhakar: *http://www.orkut.com/Profile.aspx?uid=...*

Please also let us know how many times have you emailed your highest and lowest ranked friends in the last 3 months.

C.2 Hypothesis: Role of social ties

Given a classification of ties as strong or weak, messages by strong ties of a person provide more context than weak ties, and messages by weak ties provide more completeness than strong ties.

Sample survey: Orissa

Seeing your interest in Orissa, we feel you must be concerned with the status of social developments there. Assume you had to rely on your friends to get the latest news about developments in Orissa. Please rank your following 5 friends on a scale of 1-5, based on:

(a) How well they know what kind of topics about Orissa you find interesting.

1 = Your friend does not know about your specific interests in Orissa. You have to rely on yourself to seek and understand information.

5 = Your friend knows about your interests extremely well, such that he/she can recommend useful news and explain its relevance for you.

(b) How well you feel they are aware of diverse aspects of life in Orissa.

1 = Your friend is not aware of the diverse viewpoints of different groups of people, and does not help with providing different perspectives.

5 = Your friend is very well informed about diverse perspectives and can update you with them.

1. Abhipsa: *http://www.orkut.com/Profile.aspx?uid=...*
2. Deb: *http://www.orkut.com/Profile.aspx?uid=...*
3. The Bum: *http://www.orkut.com/Profile.aspx?uid=...*
4. Akash: *http://www.orkut.com/Profile.aspx?uid=...*
5. Nick: *http://www.orkut.com/Profile.aspx?uid=...*

Sample survey: Mumbai

Seeing your interest in Mumbai, we feel you must be concerned with the transportation and sanitation infrastructure there. Assume you had to rely on your friends to get the latest news about related developments in Mumbai. Please rank your following 5 friends on a scale of 1-5, based on:

(a) How well they know what kind of topics about Mumbai you find interesting.

1 = Your friend does not know about your specific interests in Mumbai. You have to rely on yourself to seek and understand information.

5 = Your friend knows about your interests extremely well, such that he/she can recommend useful news and explain its relevance for you.

(b) How well you feel they are aware of diverse aspects of the lives of people in Mumbai.

1 = Your friend is not aware of the diverse viewpoints of different groups of people, and does not help with providing different perspectives.

5 = Your friend is very well informed about diverse perspectives and can update you with them.

Table C.1: Comparison of different scenarios for Economics

	Context	**Completeness**
Strong ties	$\mu = .80, n = 145$ $z = -1.42$	$\mu = .56, n = 145$ $z = -0.20$
Weak ties	$\mu = .26, n = 50$ $z = -5.15$	$\mu = .77, n = 50$ $z = -1.09$

Table C.2: Comparison of different scenarios for Mumbai

	Context	**Completeness**
Strong ties	$\mu = .78, n = 163$ $z = -1.82$	$\mu = .55, n = 163$ $z = -0.38$
Weak ties	$\mu = .33, n = 24$ $z = -2.94$	$\mu = .71, n = 24$ $z = -1.03$

1. varun: *http://www.orkut.com/Profile.aspx?uid=...*
2. Yayati: *http://www.orkut.com/Profile.aspx?uid=...*
3. ninad: *http://www.orkut.com/Profile.aspx?uid=...*
4. DON is Back: *http://www.orkut.com/Profile.aspx?uid=...*
5. Abir: *http://www.orkut.com/Profile.aspx?uid=...*

Results

Tables C.1, C.2, and C.3 show the results for hypothesis-1 for topics about economics, Mumbai, and books respectively.

Table C.3: Comparison of different scenarios for Books

	Context	**Completeness**
Strong ties	$\mu = .76, n = 128$ $z = -1.05$	$\mu = .51, n = 128$ $z = -0.54$
Weak ties	$\mu = .42, n = 60$ $z = -3.79$	$\mu = .64, n = 60$ $z = -2.09$

C.3 Modeling: Context and completeness of messages

Sample survey: Orissa

Seeing your membership in the community ‘A Better Odisha’, we would like to ask you two questions about the discussion titled ‘What abt filing RTI applications?’: *http://www.orkut.com/CommMsgs.aspx?cmm=...&tid=...*

1. RTI can be useful in different ways in different places. Do you feel this discussion sufficiently explains how you could use RTI to in your particular circumstances?

2. Do you feel the discussion brings in fairly diverse points of view to help you properly analyze the different choices you might have for using RTI?

Please give your ranking on a scale of 1-5 (1 = poor, 5 = excellent).

Sample survey: Economics

Seeing your membership in the community ‘Economics honours’, we would like to ask you two questions about the discussion titled ‘Post eco hons - what now?’: *http://www.orkut.com/CommMsgs.aspx?cmm=...&tid=...*

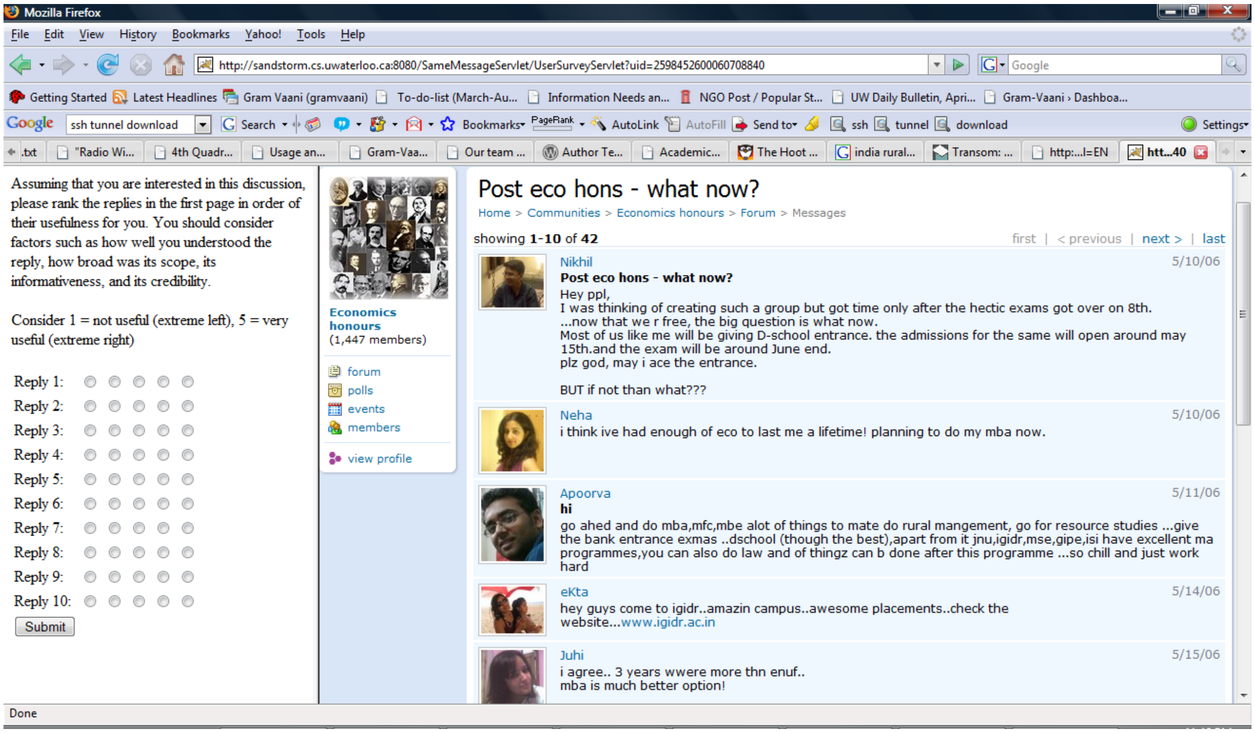
1. Each individual’s circumstances regarding professional options are likely to be different from each other. Do you feel this discussion sufficiently explains how it could be useful for you in your particular circumstances?

2. Do you feel the discussion brings in fairly diverse points of view to help you properly analyze your choices?

Please give your ranking on a scale of 1-5 (1 = poor, 5 = excellent).

C.4 Recommendation algorithm: Usefulness model

Since Orkut does not allow users to rate individual messages, we wrote a web application and invited volunteer users to survey messages for us. Fig. C.1 shows a screen-shot of our application.


Mozilla Firefox
File Edit View History Bookmarks Yahoo! Tools Help
http://sandstorm.cs.uwaterloo.ca:8080/SameMessageServlet/UserSurveyServlet?uid=2598452600060708840
Google
Getting Started Latest Headlines Gram Vaani (gramvaani) To-do-list (March-Au... Information Needs an... NGO Post / Popular St... UW Daily Bulletin, Apri... Gram-Vaani › Dashboa...
Google ssh tunnel download Search Bookmarks PageRank AutoLink AutoFill Send to ssh tunnel download Settings
.txt "Radio Wi... 4th Quadr... Usage an... Gram-Vaa... Our team ... Author Te... Academic... The Hoot ... india rural... Transom: ... http:...l=EN htt...40
Economics honours
(1,447 members)
forum
polls
events
members
view profile
Done

Assuming that you are interested in this discussion, please rank the replies in the first page in order of their usefulness for you. You should consider factors such as how well you understood the reply, how broad was its scope, its informativeness, and its credibility.

Consider 1 = not useful (extreme left), 5 = very useful (extreme right)

Reply 1: ○ ○ ○ ○ ○
Reply 2: ○ ○ ○ ○ ○
Reply 3: ○ ○ ○ ○ ○
Reply 4: ○ ○ ○ ○ ○
Reply 5: ○ ○ ○ ○ ○
Reply 6: ○ ○ ○ ○ ○
Reply 7: ○ ○ ○ ○ ○
Reply 8: ○ ○ ○ ○ ○
Reply 9: ○ ○ ○ ○ ○
Reply 10: ○ ○ ○ ○ ○

Submit

Post eco hons - what now?

Home > Communities > Economics honours > Forum > Messages

showing **1-10** of **42** first | < previous | next > | last

Nikhil 5/10/06
Post eco hons - what now?
Hey ppl,
I was thinking of creating such a group but got time only after the hectic exams got over on 8th.
...now that we r free, the big question is what now.
Most of us like me will be giving D-school entrance. the admissions for the same will open around may 15th.and the exam will be around June end.
plz god, may i ace the entrance.

BUT if not than what???

Neha 5/10/06
i think ive had enough of eco to last me a lifetime! planning to do my mba now.

Apoorva 5/11/06
hi
go ahed and do mba,mfc,mbe alot of things to mate do rural mangement, go for resource studies ...give the bank entrance exmas ..dschool (though the best),apart from it jnu,igidr,mse,gipe,isi have excellent ma programmes,you can also do law and of thingz can b done after this programme ...so chill and just work hard

eKta 5/14/06
hey guys come to igidr..amazin campus..awesome placements..check the website...www.igidr.ac.in

Juhi 5/15/06
i agree.. 3 years wwere more thn enuf..
mba is much better option!

Figure C.1: User agent survey

Appendix D

Alternative credibility assessment

We mentioned in Section 7.2 that the learned credibility model could be used to make recommendations in two ways: A content-based fashion by predicting $\mathrm{P}_{it}(\mathbf{C} \mid \mathbf{E},\mathbf{S},\mathbf{P})$ for new messages for recipient user n_i and topic t, or a collaborative-based fashion by using ratings by users observed to be correlated on **CN** or **CM**. In this appendix, we describe the content-based recommendation method. Although it does not perform as well as the collaborative-based method described in Section 7.2, it does offer us many insights into the recommendation process.

Algorithm D.1: Inference phase (ratings based)

```
Input: User $i$, Cluster $V_i$ of user $i$, Message $m$;
        Ratings $\boldsymbol{R}[\mathbf{n,m}]$ given by other users to $m$;
        Learned model for user $i$
Output: P(user $i$ will find $m$ to be credible | $\boldsymbol{R}[\mathbf{k}]$)
$p_m \leftarrow \mathrm{mean}(\boldsymbol{R}[\mathbf{j,m}].\mathbf{P}'[\mathbf{j}])_{j\in 1..n}$
$s_{im} \leftarrow \mathrm{mean}(\boldsymbol{R}[\mathbf{j,m}].\mathbf{S}'[\mathbf{i,j}])_{j\in 1..n}$
$e_{im} \leftarrow \mathrm{mean}(\boldsymbol{R}[\mathbf{j,m}].\mathbf{E}'[\mathbf{i,j}])_{j\in 1..n}$
$\mathrm{P}(c_{im}|p_{im}, s_{im}, e_{im}) \leftarrow$ MCMC on learned model for $i$
```

The algorithm works as follows. When the credibility model parameters have been learned using the EM algorithm, the model is used to directly infer the probability $\mathrm{P}_{it}(c_{ix}|e_{ix}, s_{ix}, p_x)$ that n_i will find a new message m_x to be credible. Here, the evidence variables e_{ix}, s_{ix}, p_x can be calculated in two ways:

- *Authorship*: The four types of credibilities of the message are considered to be the same as the corresponding four types of credibilities of its author with respect to n_i.
- *Ratings*: The cluster and public credibilities are calculated as the weighted mean of ratings for the message given by other users and the credibilities of these users with respect to n_i. The experienced and role based credibilities are the same as the corresponding credibilities of the message author with respect to n_i.

Given the evidence variables for the new message, and the learned Bayesian model, the probability of n_i finding the message to be credible is computed using standard belief propagation methods such as Markov-Chain-Monte-Carlo (MCMC) [123]. The outline for the ratings method is given in Algorithm-D.1.

Evaluation

We show the results for experiments to find good values of α (eqn. 3 in Section 7.2) and β (eqn. 2 in Section 7.2), and compare ratings with authorship based evidence variable computation.

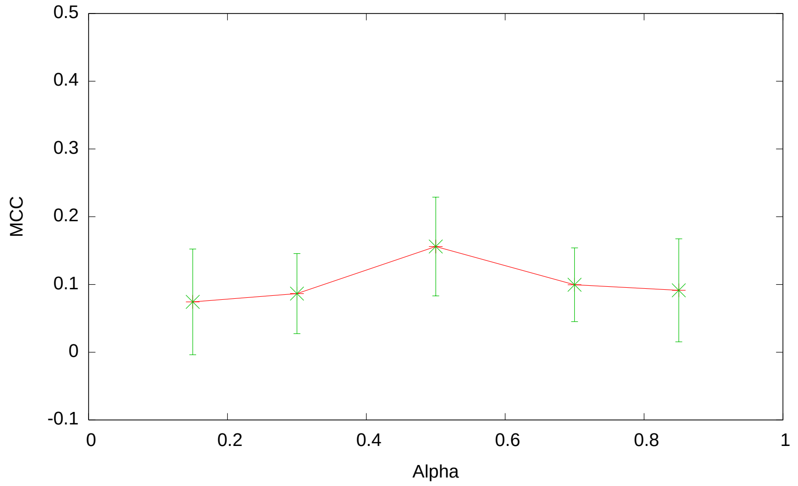


Figure D.1: Performance with different parameters

Fig. D.1 shows the mean MCC across all users for different values of α (eqn. 3) to combine the ratings and social network matrices. The best performance happens at $\alpha = 0.5$, conveying our message that all of authorship, ratings, and social networks provide valuable credibility information. All the experiments are done using ratings-based inference with $\beta = 0.85$ (eqn. 2). Larger or smaller values of β both give poorer results.

Fig. D.2 shows the TPR-FPR plot for ratings and authorship based evidence variable computation when $\alpha = 0.5$ and $\beta = 0.85$. As can be seen visually, the ratings-based method performs better than the authorship-based method. The former gives MCC = 0.156 (σ=0.073), while the latter gives MCC = 0.116 (σ=0.068). However, the authorship performance is still successful for a majority, which is encouraging. This indicates that authorship information may be used to solve the problem of cold-start for new messages that have not acquired a sufficient number of ratings. Similarly, ratings may be used to solve cold-start for new users who have not acquired sufficient credibility.

Although the content-based method for credibility computation does not perform as well as the collaborative-based method described in Section 7.2, its potential

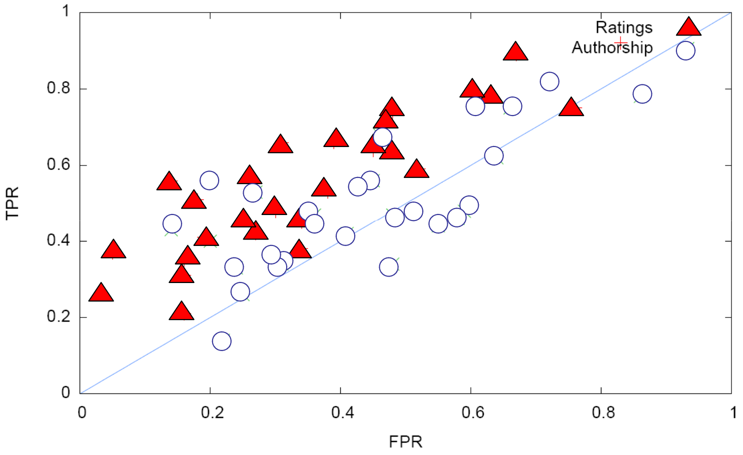


Figure D.2: Performance of Bayesian credibility model

to solve the cold-start problem in a few scenarios will be worth exploring in future work.

References

[1] Thomas F. Jones, “Poles and Zeros: Editorial,” Institute of Radio Engineers, Vol. 50, No. 12, 1962.

[2] J. Toobin, “Googles Moon Shot: The Quest for the Universal Library.,” The New Yorker, http://www.newyorker.com/reporting/2007/02/05/070205fa_fact_toobin, 2007.

[3] S. Keshav, “Why Cell Phones Will Dominate the Future Internet,” ACM CCR, April 2005.

[4] The World Bank, “The Right to Tell: The Role of Mass Media in Economic Development,” The World Bank, USA, 2002.

[5] Ian Smillie, “Mastering the Machine Revisited: Poverty, Aid, and Technology,” ITDG Publishing, UK, 2000.

[6] G. Madden and S. Savage, “Telecommunications and economic growth,” International Journal of Social Economics, Vol. 27, No. 7, 2000.

[7] The Economist, “Food: The Silent Tsunami,” http://www.economist.com/printedition/displayStory.cfm?Story_-ID=11050146, Apr 17, 2008.

[8] B. Oyelaran-Oyeyinka and C. Adeya, “Internet Access in Africa: Empirical Evidence from Kenya and Nigeria,” Telematics and Informatics, Vol. 21, No. 1, 2003.

[9] Gartner, “Gartner Says India to Have 6.9 Million Mobile and Fixed WiMAX Connections by the End of 2011,” http://www.gartner.com/it/page.jsp?id=631808, 2008.

[10] “Cellular Data Plan Comparison Chart,” http://www.jiwire.com/cellular-data-cellular-data-the-plans.htm, 2005.

[11] D. Sifry, “The State of the Live Web,” http://www.sifry.com/alerts/archives/000493.html, April 2007.

[12] J. Bryant and D. Zillman, "Media Effects: Advances in Theory and Research," Lawrence Erlbaum Associates, New Jersey, USA, 2002.

[13] Sally Jackson, "Message Effects Research: Principles of Design and Analysis," Guilford Press, New York, USA, 1992.

[14] M. Hindeman, K. Tsioutsiouliklis, and J. Johnson, "Googlearchy: How a Few Heavily Linked Sites Dominate Politics on the Web," http://www.princeton.edu/ mhindman/googlearchy–hindman.pdf, Jul 2003.

[15] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, "Topical Interests and the Mitigation of Search Engine Bias," Proc. National Academy of Sciences, Vol. 103, No. 34, Aug 2006.

[16] Marshall McLuhan, "Understanding Media," Gingko Press, Corte Madera, USA, 1964.

[17] Neil Postman, "Amusing Ourselves to Death: Public Discourse in the Age of Show Business," Penguin Group, New York, 1985.

[18] A. Jhunjhunwala, "Connecting Rural India: Towards Wealth Generation in Rural India," http://www.tenet.res.in/ Publications/Presentations/pdfs/ruralShortDec04.pdf, Dec 2004.

[19] B. Raman and K. Chebrolu, "Experiences in Using WiFi for Rural Internet in India," IEEE Communications, Jan 2007.

[20] M. Granovetter, "The Strength of Weak Ties," American Journal of Sociology, Vol. 78, No. 6, 1973.

[21] M. McPherson, L. Smith-Lovin, and J. Cook, "Birds of a Feather: Homophily in Social Networks," Annual Review of Sociology, Vol. 27, 2001.

[22] S. Hansell, "Inbox 2.0: Yahoo and Google to Turn Email into a Social Network," The New York Times, Nov 13, 2007.

[23] M. T. Hansen, "The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits," Administrative Science Quarterly, Vol. 44, 1999.

[24] T. Armstrong, O. Trescases, C. Amza, and E. Lara, "Efficient and Transparent Dynamic Content Updates for Mobile Clients," Proc. ACM/USENIX MOBISYS, Jun 2006.

[25] E. Shih, P. Bahl, and M. Sinclair, "Wake on Wireless: An Event Driven Energy Saving Strategy for Battery Operated Devices," Proc. ACM MOBICOM, 2002.

[26] Y. Mao, B, Knutsson, H. Lu, and J. Smith, "DHARMA: Distributed Home Agent for Robust Mobile Access," Proc. IEEE INFOCOM, 2005.

[27] H. Balakrishnan, V. Padmanabhan, S. Seshan, and R. Katz, "A Comparison of Mechanisms for Improving TCP Performance Over Wireless Links," In IEEE/ACM Transactions on Networking, Vol. 5, No. 6, 1997.

[28] B. Aboba, "Architectural Implications of Link Layer Indications," http://www.ietf.org/internet-drafts/draft-iab-link-indications-01.txt, Jan 2005.

[29] R. Chakravorty, A. Clark, and I. Pratt, "GPRSWeb: Optimizing the Web for GPRS Links," Proc. ACM/USENIX MOBISYS, 2003.

[30] J. Byers, M. Luby, M. Mitzenmacher, and A. Rege, "A Digital Fountain Approach to Reliable Distribution of Bulk Data," Proc. ACM SIGCOMM, 1998.

[31] M. Chan and R. Ramjee, "Improving TCP/IP Performance Over Third Generation Wireless Networks," Proc. IEEE INFOCOM, 2004.

[32] M. Demmer, E. Brewer, K. Fall, S. Jain, M. Ho, and R. Patra, "Implementing Delay Tolerant Networking," Intel Research, Berkeley, Technical Report, IRB-TR-04-020, Dec 2004.

[33] S. Jain, K. Fall, and R. Patra, "Routing in a Delay Tolerant Network," Proc. ACM SIGCOMM, 2004.

[34] S. Zhuang, K. Lai, I. Stoica, R. Katz, and S. Shenker, "Host Mobility using an Internet Indirection Infrastructure," Proc. ACM/USENIX MOBISYS, 2003.

[35] K. Fall, "A Delay-Tolerant Network for Challenged Internets," Proc. ACM SIGCOMM 2003.

[36] H. Hsieh and R. Sivakumar, "A Transport Layer Approach for Achieving Aggregate Bandwidths on Multi-homed Mobile Hosts," Proc. ACM MOBICOM, 2002.

[37] J. Scott, J. Crowcroft, P. Hui, and C. Diot, "Haggle: A Networking Architecture Designed Around Mobile Users," Conference on Wireless On-demand Network Systems and Services (WONS), 2006.

[38] M. Kozuch and M. Satyanarayanan, "Internet suspend/resume," In Workshop on Mobile Computing Systems and Applications, 2002.

[39] R. Ludwig and R. Katz, "The Eifel Algorithm : Making TCP Robust Against Spurious Retranmission," In ACM Computer Communication Review, January 2000.

[40] R. Moskowitz, P. Nikander. P. Jokela, and T. Henderson, "Host Identity Protocol," http://www.potaroo.net/ietf/ids/draft-ietf-hip-base-00.txt, 2004.

[41] J. Ott and D. Kutscher, “A Disconnection-Tolerant Transport for Drive-thru Internet Environments,” Proc. IEEE INFOCOM 2005.

[42] A. Snoeren and H. Balakrishnan, “An End-to-End Approach to Host Mobility,” Proc. ACM MOBICOM, 2000.

[43] M. Stemm and R. Katz, “Vertical Handoffs in Wireless Overlay Networks,” In Mobile Networks and Applications, Volume 3, Number 4, Pages 335-350, 1998.

[44] H. Wang, R. Katz, and J. Giese, “Policy-Enabled Handoffs across Heterogeneous Wireless Networks,” In Mobile Computing Systems and Applications, 1999.

[45] F. Zhu and J. McNair, “Optimizations for Vertical Handoff Decision Algorithms,” Proc. WCNC, 2004.

[46] T. Pering, Y. Agarwal, R. Gupta, and R. Want, “CoolSpots: Reducing Power Consumption Of Wireless Mobile Devices Using Multiple Radio Interfaces,” Proc. ACM/USENIX MOBISYS, 2006.

[47] J. Sorber, N. Banerjee, M. Corner, and S. Rollins, “Turducken: Hierarchical Power Management for Mobile Devices,” Proc. ACM/USENIX MOBISYS, 2005.

[48] A. Nicholson, Y. Chawathe, M. Chen, B. Noble, and D. Wetherall, “Improved Access Point Selection,” Proc. ACM/USENIX MOBISYS, 2006.

[49] L. Kagal, “Rei: A Policy Language for the Me-Centric Project,” http://www.hpl.hp.com/techreports/2002/HPL-2002-270.pdf, Technical report, HP Labs, 2002.

[50] A. Qureshi and J. Guttag, “Horde: Separating Network Striping Policy from Mechanism,” Proc. ACM/USENIX MOBISYS, 2005.

[51] V. Vanghi, A. Damnjanovic, and B. Vojcic, “The CDMA2000 System for Mobile Communications,” Prentice Hall, 1st Edition, pp.224, 2004.

[52] H. Velayos, “Autonomic Wireless Networking,” Doctoral thesis, TRITA-S3-LCN-0505, ISSN 1653-0837, ISRN KTH/S3/LCN/–05/05–SE, Stockholm, Sweden, May 2005.

[53] V. Zandy, and B. Miller, “Reliable Network Connections,” Proc. ACM MOBICOM 2002.

[54] D. Kutcher and J. Ott, “Service Maps for Heterogeneous Network Environments,” Proc. Mobile Data Management Conference, 2006.

[55] V. Bychkovsky, B. Hull, A. Miu, H. Balakrishnan, and S. Madden, “A Measurement Study of Vehicular Internet Access Using In Situ Wi-Fi Networks,” Proc. ACM MOBICOM, 2006.

[56] “Soekris Net4801,” http://www.soekris.com/net4801.htm, 2005.

[57] “Bytemobile,” http://www.bytemobile.com, 2006.

[58] H. Balakrishnan, K. Lakshminarayanan, S. Ratnasamy, S. Shenker, I. Stoica, and M. Walfish, ”A Layered Naming Architecture for the Internet,” Proc. ACM SIGCOMM 2004.

[59] J. Burgess, B. Gallagher, D. Jensen, and B. Levine, ”MaxProp: Routing for Vehicle-Based Disruption-Tolerant Networking,” Proc. IEEE Infocom 2006.

[60] V. Cerf, S. Burleigh, A. Hooke, L. Torgerson, R. Durst, K. Scott, K. Fall, and H. Weiss, ”Delay Tolerant Network Architecture,” Internet Draft http://www.dtnrg.org/specs/draft-irtf-dtnrg-arch-02.txt, July 2004.

[61] S. Cheshire and M. Baker, ”A Wireless Network in MosquitoNet,” In IEEE Micro, Feb 1996.

[62] S. Guo, M. Ghaderi, and S. Keshav, ”Opportunistic Scheduling in Ferry Based Networks,” Proc. WNEPT, 2006.

[63] A. Jhunjhunwalla, ”Wireless Mesh Access in Rural India,” Presentation at COMSWARE 2006, New Delhi, January 2006.

[64] E. Jones, L. Li, and P. Ward, ”Practical Routing in Delay-Tolerant Networks,” Proc. Workshop on DTN, 2005.

[65] J. Ott and D. Kutscher, ”A Disconnection-Tolerant Transport for Drive-thru Internet Environments,” Proc. IEEE INFOCOM 2005.

[66] A. Pentland, R. Fletcher, and A. Hasson, ”Daknet: Rethinking Connectivity in Developing Nations,” IEEE Computer, 37(1):78-83, 2004.

[67] C. Perkins, ”IP Mobility Suport for Ipv4,” http://www.ietf.org/rfc/rfc3344.txt, Aug 2002.

[68] S. Rhea, B. Godfrey, B. Karp, J. Kubiatowicz, S. Ratnasamy, S. Shenker, I. Stoica, and H. Yu, ”OpenDHT: A Public DHT Service and Its Uses,” Proc. ACM SIGCOMM 2005.

[69] H. Soliman, C. Catelluccia, K. Malki, and L. Bellier, ”Hierarchical Mobile IPv6 mobility management (HMIPv6),” http://www.ietf.org/internet-drafts/draft-ietf-mipshop-hmipv6-04.txt, 2004.

[70] W. Zhao, M. Ammar, and E. Zegura, ”A Message Ferrying Approach for Data Delivery in Sparse Mobile Ad Hoc Networks,” Proc. ACM MOBIHOC 2004.

[71] DTN Research Group (DTNRG), http://www.dtnrg.org/, 2006.

[72] One Laptop Per Child (OLPC), http://laptop.media.mit.edu/, 2006.

[73] C. Gentry and A. Silverberg, "Hierarchical ID-Based Cryptography," Proc. International Conference on the Theory and Application of Cryptography and Information Security, 2002.

[74] D. Boneh and M. Franklin, "Identity Based Encryption from the Weil Pairing," Proc. Crypto 2001 Lecture Notes in Computer Science, Vol 2139, Springer Verlag, 2001.

[75] B. Lampson, "Computer Security in the Real World," IEEE Computer, June 2004.

[76] C. Gentry, "Certificate based encryption and the certificate revocation problem," Proc EUROCRYPT, pp 272-293, 2003.

[77] D. Yao, N. Fazio, Y. Dodis, and A. Lysyanskasa, "ID-Based Encryption for Complex Hierarchies with Applications to Forward Security and Broadcast Encryption," Proc. ACM Conference on Computer and Communications Security, 2004.

[78] J. Mirkovic, S. Dietrich, D. Dietrich, and P. Reiher, "Internet Denial of Service: Attack and Defence Mechanisms," Prentice Hall PTR, 2005.

[79] K. Zhang and T. Kindberg, "An Authorization Infrastructure for Nomadic Computing," Proc. ACM Symposium on Access Control Models and Technologies, 2002.

[80] A. Menezes, P. Oorschot, and S. Vanstone, "Handbook of Applied Cryptography," CRC Press, 1996.

[81] R. Tiexeira, A. Shaikh, T. Griffin, and J. Rexford, "Dynamics of Hot-Potato Routing in IP Networks," Proc. ACM SIGMETRICS, 2004.

[82] K. Fall, "Identity Based Cryptosystem for Secure Delay Tolerant Networking," Manuscript, Intel Research, Berkeley, 2003.

[83] A. Kate, G. Zaverucha, and U. Hengartner, "Anonymity and Security in Delay Tolerant Networks," Proc. SecureComm, 2007.

[84] S. Ur Rehman, U. Hengartner, U. Ismail, and S. Keshav, "Practical Security for Rural Internet Kiosks," Proc. ACM Sigcomm NSDR Workshop, 2008.

[85] T. O'Reilly, "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, 2005.

[86] Jurgen Habermas, "The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society," MIT Press, USA, 1989.

[87] H. Greisdorf, "Relevance: An Interdisciplinary and Information Science Perspective," Informing Science, Special Issue on Information Science Research, Vol. 3, No. 2, 2000.

[88] F.W. Lancaster and V. Gale, "Pertinence and Relevance," Encyclopedia of Library and Information Science, CRC Press, 2000.

[89] D. Lee Howard, "Pertinence as Reflected in Personal Constructs," Journal of the American Society for Information Science, Vol. 45, No. 3, 1994.

[90] K. Maglaughlin and D. Sonnenwald, "User Perspectives on Relevance Criteria: A Comparison among Relevant, Partially Relevant, and Not-Relevant Judgements," Journal of Information Science and Technology, Vol. 53, No. 5, 2002.

[91] S. Rieh, "Judgement of Information Quality and Cognitive Authority on the Web," Journal of the American Society for Information Science and Technology, Vol. 53, No. 2, 2002.

[92] X. Zhu and S. Gauch, "Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web," Proc. SIGIR, 2000.

[93] P. Sainath, "Give us a Price, not a Package," http://www.indiatogether.org/2006/aug/psa-price.htm, Aug 2006.

[94] M. Granovetter, "The Strength of Weak Ties," American Journal of Sociology, Vol. 78, No. 6, 1973.

[95] M. McPherson, L. Smith-Lovin, and J. Cook, "Birds of a Feather: Homophily in Social Networks," Annual Review of Sociology, Vol. 27, 2001.

[96] J. Kleinberg, "The Small-World Phenomenon: An Algorithmic Perspective," In Proc. STOC, 2000.

[97] M. T. Hansen, "The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits," Administrative Science Quarterly, Vol. 44, 1999.

[98] B. Baybeck and R. Huckfeldt, "Urban Contexts, Spatially Dispersed Networks, and the Diffusion of Political Information," Political Geography, Vol. 21, 2002.

[99] D. Narayan, "Bonds and Bridges: Social Capital and Poverty," Poverty Group, World Bank, 1999.

[100] A. Krishna, "Active Social Capital: Tracing the Roots of Development and Democracy," Columbia University Press, New York, USA, 2002.

[101] "Viplav Communications: Engineering the Change," http://www.alchemists-india.com/, 2005.

[102] B. Fogg and H. Tseng, "The Elements of Computer Credibility," Proc. SIGCHI, 1999.

[103] J. Sabater and C. Sierra, "Social ReGreT, A Reputation Model Based on Social Relations," ACM SIGecom Exchanges, Vol. 3, No. 1, 2002.

[104] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proc. National Academy of Sciences, Vol. 99. No. 12, 2002.

[105] S. Wasserman and K. Faust, "Social Network Analysis," Cambridge University Press, UK, 1994.

[106] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," Proc. National Academy of Science, Vol. 101, No. 9, 2004.

[107] H. Eulau and L. Rothenberg, "Life Space and Social Networks as Political Contexts," Political Behavior, Vol. 8, 1986.

[108] L. Adamic and E. Adar, "How to Search a Social Network," Social Networks, Vol. 27, No. 3, 2005.

[109] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and S. Bhattacharjee, "Measurement and Analysis of Online Social Networks," Proc. IMC, 2007.

[110] S. Dongen, "MCL: A Cluster Algorithm for Graphs," PhD thesis, University of Utrecht, 2000.

[111] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe, "A Framework for Community Identification in Dynamic Social Networks," Proc. SIGKDD, 2007.

[112] L. Wasserman, "All of Statistics: A Concise Course in Statistical Inference," Springer-Verlag, New York, USA, 2004.

[113] M. E. J. Newman, "The Structure and Function of Complex Networks," SIAM Review, Vol. 45, No. 2, 2003.

[114] T. Valente, "Network Models for the Diffusion of Innovations," Cresskill Hampton Press, USA, 1995.

[115] A. Singhal, "Modern Information Retrieval: A Brief Overview," IEEE Data Engineering Bulletin, Vol. 24, 2001.

[116] T. Turner, M. Smith, D. Fisher, and H. Welser, "Picturing Usenet: Mapping Computer-Mediated Collective Action," Journal of Computer Mediated Communication, Vol. 10, No. 4, 2005.

[117] A. Mislove, K. Gummadi, and P. Druschel, "Exploiting Social Networks for Internet Search," Proc. Hotnets, 2006.

[118] X. Song, B. Tseng, C. Lin, and M. Sun, "Personalized Recommendation Driven by Information Flow," Proc. SIGIR, 2006.

[119] A. Das, M. Datar, A. Garg, and S. Rajaram, "Google News Personalization: Scalable Online Collaborative Filtering" Proc. WWW, 2007.

[120] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 6, 2005.

[121] M. Angermann, P. Robertson, and T. Strang, "Issues and Requirements for Bayesian Approaches in Context Aware Systems," LNCS, Springer, Berlin, Heidelberg, 2005.

[122] Gene H. Golub and Charles F. Van Loan, "Matrix Computations," John Hopkins University Press, Maryland, 1983.

[123] S. Russel and P. Norvig, "Artificial Intelligence: A Modern Approach," Pearson Education, New Jersey, USA, 2003.

[124] A. Kale, A. Karandikar, P. Kolari, A. Java, A. Joshi, and T. Finin, "Modeling Trust and Influence in the Blogosphere Using Link Polarity," Proc. ICWSM, 2007.

[125] J. Yang, J. Wang, M. Clements, J. Pouwelse, A. P. de Vries, and M. Reinders, "An Epidemic-based P2P Recommender System," Proc. SIGIR Workshop on Large Scale Distributed Systems for Information Retrieval, 2007.

[126] B. Yu and M. Singh, "Searching Social Networks," Proc. AAMAS 2003.

[127] J. Kleinberg, "Complex Networks and Decentralized Search Algorithms," Proc. ICM, 2006.

[128] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network Motifs: Simple Building Blocks of Complex Networks," Science, Vol. 298, 2002.

[129] A. Langville and C. Meyer, "A Survey of Eigenvector Methods for Web Information Retrieval," ACM-SIAM Review, 2004.

[130] P. Melville, R. Mooney, and R. Nagarajan, "Content-Boosted Collaborative Filtering for Improved Recommendations," Proc. AAAI, 2002.

[131] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," Proc. ICML, 2006.

[132] S. Brin and L. Page, "The PageRank Citation Ranking: Bringing Order to the Web," http://dbpubs.stanford.edu:8090/pub/1999-66, TechnicalReport, Stanford University, 2001.

[133] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998.

[134] K. Lerman, "Social Information Processing in News Aggregation," IEEE Internet Computing, Vol. 11, No. 6, 2007.

[135] S. Kamvar, M. Scholsser, and H. Garcia-Molina, "The EigenTrust Algorithm for Reputation Management in P2PNetworks," Proc. WWW, 2003.

[136] J. M. Pujol, R. Sanguesa, and J. Delgado, "Extracting Reputation in Multi Agent Systems by Means of Social Network Topology," Proc. AAMAS, 2002.

[137] K. Walsh and E. Gun Sirer, "Experience with an Object Reputation System for Peer-to-Peer Filesharing," Proc. USENIX NSDI, 2006.

[138] J. Zhang and R. Cohen, "A Personalized Approach to Address Unfair Ratings in Multiagent Reputation Systems," Proc. AAMAS Workshop on Trust in Agent Societies, 2006.

[139] A. Whitby, A. Josang, and J. Indulska, "Filtering Out Unfair Ratings in Bayesian Reputation Systems," Icfain Journal of Management Research, 2005.

[140] U. Kuter and J. Golbeck, "SUNNY: A New Algorithm for Trust Inference in Social Networks Using Probabilistic Confidence Models," Proc. AAAI, 2007.

[141] T. Huynh, N. Jennings, and N. Shadbolt, "FIRE: An Integrated Trust and Reputation Model for Open Multi-agent Systems," Proc. European Conference on Artificial Intelligence, 2004.

[142] P. Kolari, T. Finin, K. Lyons, Y. Yesha, Y. Yesha, S. Perelgut, and J. Hawkins, "On the Structure, Properties, and Utility of InternalCorporate Blogs," Proc. ICWSM, 2007.

[143] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of Trust and Distrust," Proc. WWW, 2004.

[144] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub, "Exploiting the Block Structure of the Web for Computing Pagerank," Technical report, Stanford University, http://www-nlp.stanford.edu/pubs/blockrank.pdf, 2003.

[145] L. Kaelbling, M.Littman, and A. Moore, "Reinforcement Learning: A Survey," http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume4/kaelbling96a-html/rl-survey.html, 1996.

[146] I. Rose, R. Murty, P. Pietzuch, J. Ledlie, M. Roussopoulos, and M. Welsh, “Cobra: Content-based Filtering and Aggregation of Blogs and RSS Feeds,” Proc. USENIX NSDI, 2007.

[147] V. Ramasubramanian, R. Peterson, and G. Sirer, “Corona: A High Performance Publish-Subscribe System for the World Wide Web,” Proc. USENIX NSDI, 2006.

[148] L. Carbera, M. Jones, and M. Theimer, “Herald: Achieving a Global Event Notification Service,” Proc. Workshop on Hot Topics in Operating Systems, 2001.

[149] F. Fabret, H. A. Jacobsen, F. Llirbat, J. Pereira, and K. A. Ross, “Filtering Algorithms and Implementation for Very Fast Publish/Subscribe Systems,” Proc. ACM SIGMOD, 2001.

[150] M. Rattigan, M. Maier, and D. Jensen, “Using Structure Indices for Efficient Approximation of Network Properties,” Proc. ACM SIGKDD, 2006.

[151] C. Intanagonwiwat, R. Govindan and D. Estrin, “Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks,” Proc. ACM MOBICOM, 2000.

[152] J. Travers and S. Milgram, “An Experimental Study of the Small World Problem,” Sociometry, Vol. 32, No. 425, 1969.

[153] P. Dodds, R. Muhamad, D. Watts, “An Experimental Study of Search in Global Social Networks,” Science, Vol. 301, No. 5634, 2003.

[154] D. Watts, P. Dodds, and M. Newman, “Identity and Search in Social Networks,” Science, Vol. 296, No. 5571, 2002.

[155] F. Dabek, R. Cox, F. Kaashoek, and R. Morris, “Vivaldi: A Decentralized Network Coordinate System,” Proc. ACM SIGCOMM, 2004.

[156] I. Bhattacharya and L. Getoor, “A Latent Dirichlet Model for Unsupervised Entity Resolution,” SIAM-SDM, 2006.

[157] A. Carzaniga, M. Rutherford, and A. Wolf, “A Routing Scheme for Content-Based Networking,” Proc. IEEE Infocom, 2004.

[158] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis,” American Society for Information Science, 1990.

[159] K. Sripanidkulchai, B. Maggs, and H. Zhang, “Efficient Content Location Using Interest-Based Locality in Peer-to-Peer Systems,” Proc. IEEE Infocom, 2003.

[160] David Liben-Nowell, “An Algorithmic Approach to Social Networks,” PhD thesis, MIT Computer Science and Artificial Intelligence Laboratory, 2005.

[161] E.F. Schumacher, “Small is Beautiful: Economics as if People Mattered,” Hartley and Marks, Vancouver, Canada, 1973.

[162] BBC News, “Design Revamp for $100 Laptop,” http://news.bbc.co.uk/2/hi/technology/7411904.stm, 2008.

[163] A. Garai and B. Shadrach, “Taking ICT to Every Indian Village,” One World South Asia, New Delhi, India, 2006.

[164] J. Kendall and N. Singh, “Internet Kiosks in Rural India: What Influences Success,” Working Paper, NET Institute, 06-05, 2006.

[165] A. Kumar, N. Rajput, D. Chakraborty, S. Agarwal, and A. Nanavati, “WWTW: The World Wide Telecom Web,” Proc. ACM SIGCOMM Workshop on Networked Systems for Developing Regions, 2007.

[166] R. Patra, S. Nedevschi, S. Surana, A. Sheth, L. Subramanian, and E. Brewer, “WiLDNet: Design and Implementation of High Performance WiFi Based Long Distance Networks,” Proc. USENIX NSDI, 2007.

[167] S. Surana, R. Patra, S. Nedevschi, M. Ramos, L. Subramanian, Y. Ben-David, and E. Brewer, “Beyond Pilots: Keeping Rural Wireless Networks Alive,” Proc. USENIX NSDI, 2008.

[168] George Orwell, “1984,” Signet Classic, New York, USA, 1949.

[169] D. Hadaller, S. Keshav, T. Brecht, and S. Agarwal, “Vehicular Opportunistic Communication Under the Microscope,” Proc. ACM MOBISYS, 2007.

[170] M.A. Zaharia and S. Keshav, “Fast and Optimal Scheduling Over Multiple Network Interfaces,” Technical Report, University of Waterloo, CS-2007-36, 2007.

[171] V. Srinivasan, M. Motani, and W. Ooi, “Analysis and Implications of Student Contact Patterns Derived from Campus Schedules,” Proc. ACM MOBICOM, 2006.

[172] H. Raiffa and R. Schaifer, “Applied Statistical Decision Theory,” Harvard Business School Publications, USA, 1961.

[173] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava, “Participatory Sensing,” Proc. SenSys Workshop on the World-Sensor-Web, 2006.

[174] J. Zhang, M. Ackerman, and L. Adamic, “Expertise Networks in Online Communities: Structure and Algorithms,” Proc. WWW, 2007.

[175] T. Lento, H. Welser, L. Gu, and M. Smith, "The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System," Proc. WWW Workshop on the Weblogging Ecosystem, 2006.

[176] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Laws of Graph Evolution: Densification and Shrinking Diameters," ACM Transactions on Knowledge Discovery from Data, 2007.

[177] R. Albert, and A. Barabasi, "Statistical Mechanics of Complex Networks," Review of Modern Physics, Vol. 74, 2002.

[178] J. Kurose and K. Ross, "Computer Networking," Addison Wesley, 3rd Edition, pg. 271, 2004.

[179] B. Hogan, J. Carrasco, and B. Wellman, "Visualizing Personal Networks: Working with Participant-Aided Sociograms," Field Methods, Vol. 19, No, 2, 2007.

[180] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis, "Tastes, Ties, and Time: A New (Cultural, Multiplex, and Longitudinal) Social Network Dataset Using Facebook.com," Under review, 2007.

[181] Daniel Riffe, Stephen Lacy, and Frederick Fico, "Analyzing Media Messages: Using Quantitative Content Analysis in Research," Lawrence Erlbaum Associates, New Jersey, USA, 2005.

[182] T. Gruber, "Ontology," To appear in *Encyclopedia of Database Systems*, L. Liu and M. Tamer zsu, Springer Verlag, USA, 2008.

[183] Norman Bradburn, Seymour Sudman, and Brian Wansink, "Asking Questions: The Definitive Guide to Questionnaire Design for Market Research, Political Polls, and Social and Health Questionnaires," John Wiley and Sons, San Franciso, USA, 2004.

[184] A. Seth, P. Darragh, S. Liang, Y. Lin, and S. Keshav, "An Architecture for Tetherless Communication," http://blizzard.cs.uwaterloo.ca/keshav/home/Papers/data/05/tca.pdf, Manuscript, University of Waterloo, July 2005.

[185] A. Seth, S. Bhattacharya, and S. Keshav, "Opportunistic Communication Over Heterogeneous Access Networks," http://www.cs.uwaterloo.ca/ a3seth/ocmptech.pdf, Technical report, Sprint Labs, CA, April 2005.

[186] A. Seth, "The Use of Mobile Devices as Sensors for Efficient Wireless Monitoring and Resource Utilization," Manuscript, University of Waterloo, 2006

[187] A. Seth, D. Kroeker, M. Zaharia, S. Guo, and S. Keshav, “Low-cost Communication for Rural Internet Kiosks Using Mechanical Backhaul,” Proc. ACM MOBICOM, 2006.

[188] A. Seth, U. Hengartner, and S. Keshav, ”Practical Security for Disconnected Nodes,” Manuscript, University of Waterloo, 2005.

[189] A. Seth and S. Keshav, ”Practical Security for Disconnected Nodes,” Proc. ICNP Workshop on Secure Network Protocols, 2005.

[190] A. Seth, “An Infrastructure for Participatory Media,” Proc. AAAI Workshop on Recommender Systems, 2007.

[191] A. Seth, “Understanding Participatory Media Using Social Networks,” Technical Report CS-2007-47, University of Waterloo, 2007.

[192] A. Seth and J. Zhang, “A Social Network Based Approach to Personalized Recommendation of Participatory Media Content,” Proc. ICWSM, 2008.

[193] A. Seth, J. Zhang, and R. Cohen, “A Subjective Credibility Model for Participatory Media,” Proc. AAAI Workshop on Recommender Systems, 2008.

[194] A. Seth, J. Zhang, and R. Cohen, “A Multi-Disciplinary Approach for Recommending Weblog Messages,” Proc. AAAI Workshop on Enhanced Messaging, 2008.

[195] A. Seth, “Design of a Social Network Based Recommender System for Participatory Media Content,” Manuscript, University of Waterloo, 2008.

[196] A. Seth, “A Social Network Based Approach to the Evolutionary Analysis of Participatory Messages,” Proc. Web Intelligence, 2008.