Understanding Participatory Media Using Social Networks

Aaditeshwar Seth – a3seth@uwaterloo.ca

School of Computer Science
University of Waterloo, ON, Canada

Technical report: CS-2007-47

January 14, 2008

Contents

Abstract

Given the rapid growth of participatory media content such as blogs, user created videos, and podcasts, there is a need to understand and answer the following kind of questions: What is participatory media good for? Does it improve the effectiveness of news media? If so, then why? What are the underlying processes of human communication that can explain how people perceive and interpret the information they gain from participatory media? How can this understanding be used to design better information search and recommendation systems for the Internet? We try to answer these questions by examining participatory media through models built using social networks of people.

We first give a few examples of participatory media which indicate that participation by people in online public discourse through blogs and forums can indeed improve the effectiveness of news media: it helps people gain a better understanding of topics discussed in mass media, and presents people with diverse viewpoints to avoid media bias. We then use social networks to propose a model for the underlying processes of communication to answer why and how participatory media would improve the effectiveness of news media. An investigation methodology is then outlined for validation of the model, and validation results from user-surveys and measurements done on an online social networking website are presented. Our results are positive and indicate that a social network based framework offers a plausible approach to the design of useful applications for participatory media.

Chapter 1

Introduction

The news media is witnessing two interesting trends in this decade:

- People rely almost twice more on Internet portals and news websites to get their daily news, than on traditional sources such as television and newspapers. The ratio is projected to increase further [1].
- More than 1.4 million blog posts are written each day, and a large percentage of these posts are on political topics or other elements of news [2]. The number of people reading blogs on a regular basis has also been continuously increasing over the years [3].

These trends indicate a shift from traditional methods of news access and how people discuss news. We are interested in the implications of these trends on the effectiveness of news media. To avoid ambiguity, by *news media* we refer to news and opinions about current topics that are presented through channels of newspapers, television, radio, and the Internet. We consider news media to be positively *effective* if they are simple and easily understandable by the people, and if they provide the people with complete information to give them an unbiased viewpoint. We use *participatory media* to refer to forms of media such as blogs and online forums, which allow people to participate in news discussions. The question we want to answer is whether participatory media can improve media effectiveness, and if so, then under what conditions. We next give a few examples of participatory media, and show why it is worthwhile to research this question. In the next chapter, we describe a framework for this research, and then outline a methodology to conduct the research.

1.1 Examples

One of the main application of news media is in politics, and there is documented anecdotal evidence [4,5] which indicates that participatory media enabled by the Internet indeed improves *media effectiveness*. We next give a few more examples of participatory media in other political and non-political topics to understand it better. Some of these examples may not necessarily form a part of news, but are nonetheless useful to demonstrate various aspects of participatory media.

1.1.1 News websites

Most news websites now allow people to comment on online published articles. Consider the following BBC News article about the recent Emergency declared in Pakistan, dated November 4^{th} 2007 and titled *Musharraf defends emergency rule*. 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]. Following are two comments on the same article.

- "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..."
- "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..."

Both these comments explored aspects of the event that had not been considered in the original article. In addition, the insights from the first comment are likely to be useful for other people in similar circumstances, and could spur some corrective actions on their part. Furthermore, both the comments highlight considerably diverse viewpoints in which people interpret the same event. This shows that comments by people can improve the effectiveness of the news article.

1.1.2 Book reviews

Websites such as Amazon.Com allow people to post book reviews. Consider the following reviews given for a book titled *Pattern Recognition and Machine Learning (Information Science and Statistics)*, by *Christopher M. Bishop* [2].

- "Excellent book for pattern analysis and classification! It begins with basic data curve fitting, linear classification models and ends with combining models (tree-based models, graphical models, etc). Contains great number of examples and exercises. Very good introductory for beginners in pattern analysis, excellent companion for academics and researchers."
- "I must point out that the book is very math heavy. Inspite of my considerable background in the area of neural networks and statistics, I still was struggling with the equations. This is certainly not the book that can teach one things from the ground up, and thats why I would give it only 3 stars. I am new to kernels, and I am finding the relevant chapters difficult and confusing. This book wont be very useful if all you want to do is write machine learning code. The intended audience for this book I guess are PhD students/researchers who are working with the math related aspects of machine learning. Undergraduates or people with little exposure to machine learning will have a hard time with this book. But that said, time spent in struggling with the contents of this book will certainly pay-off, not instantly though."

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

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

The book description given by the publishers only described the contents of the book, but the reviews seem to provide additional information about who is the intended audience for the book. However, both the reviewers appear to give opposite views, the first reviewer claiming that the book is good for beginners, and the second reviewer claiming that it is not. This seems to be contradictory, but the reviews are more comprehensible if the academic backgrounds of both the reviewers are considered. It is likely that the first reviewer has a much better mathematical background than the second reviewer. This shows that ad-hoc presentation of reviews can cause confusion, but a closer examination of the reviews and reviewers can avoid confusion and improve the effectiveness of the book description.

1.1.3 Question-Answer forums

Most social networking websites allow people to create discussion forums and exchange messages with each other. Consider a discussion in a forum on Orkut about graduate school options for students studying economics in India [3]. Two replies are given below.

- "Hey guys, I m also doing my M.A. in Econ from University of Akron, OH, USA. They do offer a number of assistantships which give u complete tuition waiver and also provide stipend. They do accept 15 years of education... So give it a shot... Cheers"
- "Hi everybody... I just graduated in eco and my aggregate is 59%. I am planning to drop a year for Delhi School of Economics... Wanted to know if they give weightage to your graduation aggregate or not? Or is the entrance test the main criteria for admission?"

The replies propose two different options that students can consider, namely, to drop a year and prepare well for the entrance examinations, or to consider applying to schools abroad. Unlike the previous example of book reviews, the replies in this example do not contradict each other, but their relative effectiveness still depends upon their relevance for different readers. For example, it may depend upon whether the reader has sufficient funds to apply abroad, or whether the reader can afford to drop a year for further preparation. On the whole however, participation by people does improve the effectiveness of getting useful answers, because it is unlikely that such information could have alternatively been found on a webpage somewhere.

1.2 Research novelty and relevance

The examples given here show that participation by users can indeed improve the effectiveness of gaining information, but only under certain conditions. Given the benefits of participatory media, we next explain why research on participatory media is novel and relevant.

The research is novel because participatory media presents a communication paradigm unprecedented so far. This kind of communication is unique because it simultaneously provides opportunities to people for *one-to-many* communication with other people, along with *bidirectional* information exchange among them. This enables mutual and collaborative information sharing among people. Such tools that combined bidirectional and one-to-many communication were not available at such large scales earlier. For example, prior mass media such as newspapers, television, and radio could broadcast information to many people simultaneously, but did not have any reverse communication channel. Similarly, other forms of media such as books did not provide a reverse channel either.

[3] http://www.orkut.com/CommMsgs.aspx?cmm=13213427&tid=2463642465620523221

Table 1.1: Characterization of media delivery mechanisms

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

Letters, on the other hand, provided a bi-directional communication channel, but did not facilitate one-to-many communication. This is shown in Table 1.1, indicating that participatory media is different from traditional forms of media.

The research is relevant because it will help improve the design of information search and recommendation services for the Internet, such as Google.Com and Digg.Com. These services assist people to manage the glut of information created today. Given that over 1.4 million blog posts are written daily [2], successful identification of useful blogs and discussions becomes imperative to avoid overloading users with too much information [7–9]. The second reason to improve the design of information services is the growing influence of participatory media in shaping public opinion. For example, the recent *Blog Action Day* on October 15th focused on the environment and witnessed large scale participation by over 20,000 blogs reads by more than 14 millions readers [10]. Similarly, blog websites such as OhMyNews.Com were instrumental in determining the results of the 2002 S. Korea presidential elections [4, 5]. The news media has always influenced policy formulation in democracies [11], and the role of participatory media is getting harder to deny. Since information services such as Google.Com are now fast turning into the lens through which people access this information, it therefore becomes important that these services should avoid bias or the publicity of false information to form more efficient democratic societies. However, there is evidence that the algorithms used by Google.Com often tend to bias results towards one or the other viewpoint [12], although this study was contradicted subsequently [13]. Given the importance of presenting unbiased information, and the uncertainty in the success of present systems to ensure this, it becomes important to understand the processes of participatory media more closely so that the insights can be used to improve the design of automated information services. This project suggests a timely research agenda to do so.

Chapter 2

Research framework

In the previous chapter, we showed through examples that participatory media can improve media effectiveness, but only under certain conditions. To study this further, it is useful to develop a model so that precise questions can be asked to gain insights into the topic. The model should be simple enough to restrict the scope of the questions within testable boundaries, yet flexible enough so that complications can be added to it if necessary. In addition, a model grounded in known theory is likely to be better because it can build upon the insights gained in previous research. We next describe a few such theories, and then develop a model based on these theories.

2.1 Theoretical grounding: Why

Why should we expect participatory media to improve media effectiveness at all? The answer is provided by Habermas's theory about the transformation of the public sphere, and his theory of communicative action.

2.1.1 Habermas: Structural transformation of the public sphere

In 1962, Habermas proposed a theory on the interaction between the *private* and *public* spheres in society [14] (discussed in greater details in Appendix A). He suggests that traditionally the private sphere of families and friends interacted through personal letters, *salon* discussions, and home gatherings, to create public opinion and promote a humanitarian perspective in public services provided by the governments in social welfare states. However, the creation of the modern bourgeoise society transformed the spheres when the centralized press replaced the *salons*, and began manipulating public opinion.

Interestingly, today's participatory media can be considered as reconstructing the traditional scenario of personal letters and decentralized *salon* discussions [4]. For example, many blogs are personal narratives of people, akin to letters, which could reinforce the humanitarian perspective in public opinion. Similarly, websites such as LiveJournal.Com also freely allow people to write blogs and share them with their friends [15]. Some other blogs actively debate public policy and have thousands of readers, making them similar to large scale *salon* discussions, or even comparable to

[4] In fact, *Salon.Com* is one of the most popular blog websites having many eminent scientists, civil society workers, and artists as its contributing members.

mass media. For example, one of most popular blogs, BoingBoing.Net, gets 7.5 million page-views a month, and has even started producing television shows lately [16]! If Habermas's theory is to be believed, then such personal and opinion blogs would help reinvigorate the role of the private sphere in shaping public opinion. Therefore, participation by people is likely to improve media effectiveness by questioning and diversifying the information delivered by various forms of mass media.

2.1.2 Habermas: Theory of communicative action

The role of participation to help people gain clarity and better understanding of different topics, is also corroborated by another theory of Habermas, that of communicative action, proposed in 1981 [17] (described in greater detail in Appendix B) . Habermas uses the two-level Marxian model to differentiate between the *lifeworld* for social reproduction of values and culture, and the *system* for materialistic reproduction of economic goods and services. Communication in the lifeworld, called *communicative action*, is always aimed to help people reach a common understanding about some topic. Habermas claims that communicative action between agents can succeed in achieving this understanding by following a set of discourse rules for arguments put forth by the agents. For example, one of the rules states that agents are allowed to freely challenge the arguments given by other agents, and who then have to justify the arguments they put forth. It is claimed that discourses obeying such rules eventually lead to a common understanding among the agents.

The argumentative nature and requirement for answerability in participatory media discourses makes them very similar to communicative action, and can hence be expected to lead to increased understanding for people [18]. Habermas also says that the success of communicative action depends upon the linguistic and cultural similarities between the agents, which helps them understand each other easily. The same indicator also exists in participatory media because blogs are often exchanged among mutual friends, who can be expected to share the same cultural backgrounds enabling them to understand each other more easily.

2.2 Theoretical grounding: How

Although Habermas's theories help explain why participatory media would improve media effectiveness, they do not say anything about the actual mechanisms through which people gain diverse viewpoints or more understanding. They are too abstract and extra levels of detail are required to convert the theories into operational models. Gerbner's general model of communication presents a starting point, which we later enhance using theories about social networks.

2.2.1 Gerbner: General model of communication

Gerbner proposed a general model of communication in 1956, to address many shortcomings of previous models by Shannon, Lasswell, and Schramm (described in greater detail in Appendix [19,20]). Here, communication is defined as an exchange of messages in the form of words spoken by a person, or news articles written by a news agency, etc. According to the model, an observer interprets an event or a message about the event according to his own perceptions, codifies his observations in another message, and transmits it to some other recipient. This recipient may transmit the message further to another recipient, and so on. However, at each step, the perceptions of the observer influence the message she creates, and this affects the way in which the original event is perceived by the next recipient. The model also allows for an arbitrary placement of multiple observer-recipient

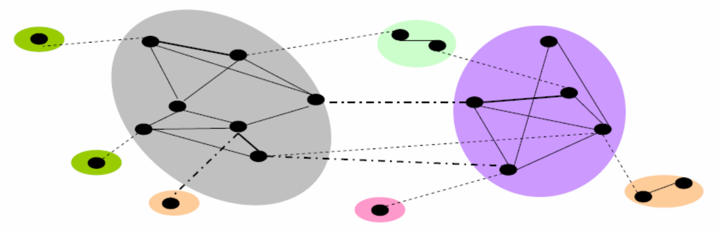

Figure 2.1: Strong and Weak Links

links in the form of a graph. This means that a single recipient who receives messages from different observers, tends to get a more diverse perspective about the original event from these observers.

Participatory media can be easily considered in terms of Gerbner's model: Bloggers write about events from their own perspectives, and this affects the interpretation of the event drawn by people reading their blog posts. Therefore, a person reading multiple blogs tends to get a more diverse perspective. This model can also explain some of the examples given in the previous chapter. Thus, people who read comments on the BBC news article will get a more complete picture about the event because the comments would have analyzed the event from multiple viewpoints. Similar gains in media effectiveness can be explained for book reviews and questions about career options. However, the model cannot describe the conditions under which participatory media will improve media effectiveness, or predict in advance which comments will be more useful than others. Granovetter's hypothesis of the *strength-of-weak-ties* in social networks becomes helpful here.

2.2.2 Granovetter: Strength of weak ties

Gerbner's model laid the foundation of communication happening along links between people. Since people are embedded in social networks and communicate among themselves, it is therefore natural to consider Gerbner's model laid out on a social network of people. But to do so, it is important to first understand certain properties of social networks.

Granovetter was among the first to show that social networks have inherent structures to them, which can explain various forms of information flow. In his famous *strength-of-weak-ties* hypothesis proposed in 1973 [27], he stated 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. 2.1. Whereas strong ties are typically constituted of close friends, weak ties are constituted of acquaintances or remote colleagues. 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 [28].

In the context of Habermas's theories and Gerbner's model, this indicates that communication among strong ties is likely to help the people understand each other efficiently, because they would share the same context with each other. However, communication across weak ties is likely to help achieve greater diversity in forming public opinion or interpreting events. In 2001, McPherson corroborated the hypothesis by showing that social networks tend to have homophily in clustering because people similar in terms of geographical location, race, ethnicity, income status, etc, tend to be clustered together [29]. This gives further rationale to the arguments because people with similar backgrounds can be expected to help more in simplifying and understanding various issues.

With reference to the examples from the previous chapter, book reviews or forum replies or comments from weak ties of a recipient are likely to provide diverse perspectives to a recipient. Messages from strong ties of a recipient are likely to be more simple and understandable though, because they would express similar perspectives in interpreting the event. Thus, in the case of the book review example, the first review is more comprehensible to other people who also have an advanced mathematics background. This is explained by the homophily argument: People strongly connected to the reviewer are likely to have a similar background. Similarly, the second review is more suited to people only looking for an introduction to the subject, some of whom are likely to be strongly connected to the second reviewer. Therefore, considering the relative position of the message author with respect to a message recipient in the social network, can explain which messages will be useful and for what reasons.

A similar aspect was also explored by Hansen in 1992 [30]. Hansen 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. This is agreement with the insights drawn above.

The social network models are however still too simple, because they are unable to explain how to precisely categorize ties as strong or weak, depending upon the message under consideration. Should a person considered as a strong tie to receive political information, also be considered as a strong tie to receive book reviews and career options? We resolve this issue by developing a more comprehensive model for participatory media.

2.3 Model for participatory media

It is useful to define two theoretical constructs at this point, called context and completeness, to develop a precise model.

2.3.1 Context

The word context is used in many ways. We henceforth use it only to denote a *set of circumstances considered in a communication task.* For the purposes of the model, the communication task is only the reading or writing of a message by a person. It is assumed that a message always refers to a real phenomenon, called an event, preserving the same terminology as that used by Gerbner. 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 writer for describing the event.* This is made more clear by referring to the examples given in the previous chapter:

- **News comments**: The context considered by the first person was his job prospects. His friends or batchmates are likely to consider the same context when reading the comment because they may find themselves to be in similar circumstances.
- **Book reviews**: The context considered by both the reviewers was their respective mathematical backgrounds, which was different in both the cases. A review would be more useful to those people who have the same background as the reviewer, that is, those people who share the same context.
- **Career options**: The context considered by both the repliers was different. The first suggestion about studying abroad possibly assumed a knowledge of good English or access to

> sufficient funds to send applications. The second reply about dropping a year is likely to have considered factors such as finding a temporary job to support himself while he prepares for the next year's entrance examinations. Different readers find different replies to be useful, depending the their own set of circumstances that would influence their decision.

Therefore, the following variables 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 writer c_s. I next present a few definitions and assumptions to build the model.

Contextual message: A message written in the same or similar context as that considered by the recipient, that is, $c_r \sim c_s$. Note that c_r and c_s are both instances of a set of circumstances. Hence, it is also possible to consider $c_r \bigcap c_s$ as the degree to which a message is contextual to a recipient. I sometimes refer to contextual messages as *context providing messages*.

Related messages: Messages about the same event e.

Assumption 1: Contextual messages are more simple and easily understandable to recipients. Thus, understanding about an event is assumed to be an outcome of reading contextual messages describing the event. If u denotes a measure of the degree to which a message m is understandable to recipient r, then the assumption states that $u \propto |c_r \bigcap c_s|$.

Assumption 2: For any particular event, people within a certain social network neighborhood of each other, have the same context for reading or writing a message about the event. Here, the social network neighborhood refers to people "close" to each other in a social network graph of personal relationships. A precise definition of "closeness" will be given later, but it is assumed that such social network neighborhoods exists for all events. However, no claims are made as yet on how to find them.

Contextual boundary: According to the assumption above: given an event e and a message recipient r, the contextual boundary b is defined as the social network neighborhood of r within which people have the same context for reading or writing a message about e.

Adjacent contextual boundaries: It is implicit in the assumption given above, that people outside a contextual boundary have a different context for reading or writing a message about the event. Therefore, given an event e, a message recipient r, and the contextual boundary b for r, the *immediate* neighboring boundaries in which people have a different context than r are together referred to as the *adjacent contextual boundaries* b'. Note that only the immediately adjacent boundaries are considered in the definition.

Hypothesis for context: It is now possible to combine the assumptions into a single testable hypothesis. For any event e and message recipient r, there exists a contextual boundary around r, such that related messages about e written by people within the boundary (set of s senders) tend to be more understandable for r than messages written by people in adjacent boundaries (set of s' senders). Here, the "tends to be more" clause denotes the probability (or rate) of writing contextual messages within and in adjacent contextual boundaries.

Only adjacent contextual boundaries are considered because the context for boundaries beyond these may or may not overlap with the context for the boundary about event e for recipient r. Experiments can now be designed to test this hypothesis in a straightforward manner. The first task is to develop an algorithm such that given an event e and a recipient r, the algorithm can construct a contextual boundary around r. Then, measure the outcome of the theoretical construct of context in terms of an observable factor which demonstrates the understanding u gained by recipient r from messages written by senders s and s'. The hypothesis will not be falsified if for a

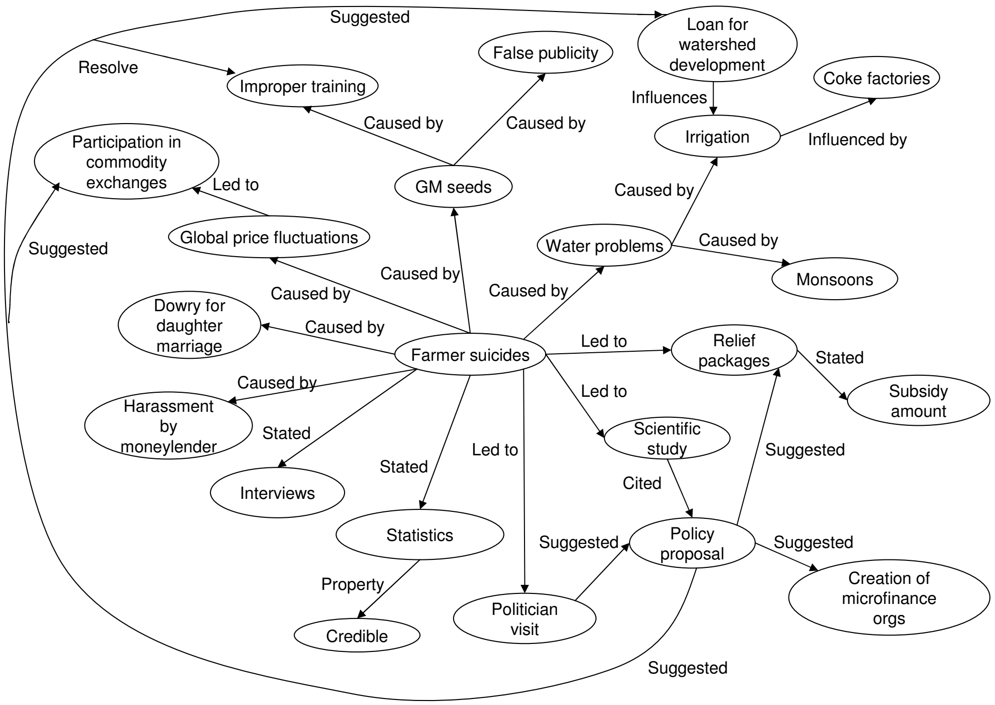

Figure 2.2: Sample ontology about farmer suicides in India

statistically significant fraction of recipients, $E_{s \in b}[u] > E_{s' \in b'}[u]$; that is, the average understanding gained from messages written by senders within the contextual boundaries of recipients, is more than the average understanding gained from messages written by senders in adjacent contextual boundaries. A method to estimate $E[u]$ will be described later in the next chapter.

2.3.2 Completeness

We use completeness to denote the *degree to which relevant aspects of an event are analyzed*. Therefore, it is possible to define the completeness of a message as the degree to which relevant aspects of the event are discussed in the message. Referring to the examples given earlier:

- **News comments**: The comments increased the completeness of the original news article by adding more perspectives about the effects of the Emergency.
- **Book reviews**: The reviews increased the completeness of the book description by adding information about the intended audience for the book.
- **Career options**: The replies increased the completeness for information about options that can be considered by different students.

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 $a_m \subseteq a_e$.

This can be explained more precisely through Fig. 2.2, which shows an ontology of relationships between various aspects relevant to an event about suicides by cotton farmers in India [31]. 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 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, t denotes the completeness of m, and $t = \frac{|a_m|}{|a_e|}$. This can be easily generalized into the average completeness $E[t]$ of a set of messages M as well: $E[t] = \frac{|a_M = \bigcup a_m|}{|a_e|.|M|}$. The exact form of $|a_m|$ is left undefined for now. It can simply be the number of ovals in Fig. 2.2 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.

The assumption is that people within the same contextual boundary tend to focus on the same or similar aspects of an event. Hence, messages from adjacent contextual boundaries provide more completeness. This assumption is used to state a testable hypothesis for completeness.

Hypothesis for completeness: For any event e and recipient r, the average completeness provided a set of related messages about e written by people within the contextual boundary for r (set of s senders), tends to be less than the average completeness provided by messages written by people in adjacent boundaries (set of s' senders) and within the contextual boundary for r. As before, the "tends to be more" clause denotes the probability or rate at which a set of messages written within various boundaries provides completeness.

This hypothesis relies on the same algorithm mentioned earlier to find contextual boundaries. Experiments can now be designed to test the hypothesis in a straightforward manner, by measuring the outcome of the theoretical construct of completeness for recipient r in terms of the measurable factor t derived through content analysis of messages written by senders s and s'. The hypothesis will not be falsified if for a statistically significant fraction of recipients, $E_{s \in b}[t] < E_{s \in b \bigcup s' \in b'}[t]$; that is, the average completeness of messages written by senders within the contextual boundaries of recipients, is less than the average completeness of messages written by senders within the same and adjacent contextual boundaries of recipients.

To summarize, we assumed that there is an algorithm which can identify contextual boundaries for an event e and recipient r. Then we hypothesized that messages written by people within the contextual boundary, are more simple and understandable by r than messages written by people in adjacent contextual boundaries. However, messages written by people in the same contextual boundary tend to focus on the same aspects of an event. Hence, messages from adjacent boundaries should be considered to gain more completeness. There are now two tasks remaining. First, to find a suitable algorithm that can identify contextual boundaries. We will use insights on social networks by Granovetter and Hansen to develop such an algorithm. Second, from the explanation of completeness given above, if messages written by people outside the contextual boundary are less understandable for r even though they provide more completeness, it is questionable whether they can improve media effectiveness for r. I resolve this confusion in my final presentation of the model by including a temporal dimension for the flow of information on social networks. These remaining tasks are described next.

2.3.3 Contextual boundaries

Before describing the procedure to identify contextual boundaries, we first introduce the notion of a topic specific social network.

Topic specific social networks: Recall that a contextual boundary was defined for a given event e and recipient r. Notice that an event which is significant for r may not be significant for other members in the social network of r. Referring to the examples given earlier, the event of the Emergency would probably be significant only for members who are from Pakistan, or who have relatives or businesses in Pakistan. The social network of r may however consist of many more people, and these people need not be considered in the contextual boundary at all. Similarly, considering a book purchase as an event, members who are not interested in the subject of machine learning need not be considered in the inference of contextual boundaries for the event. Therefore, we introduce the notion of *topic specific social networks*, which consists of the induced subgraph of only those people and links between people who are interested in the same topic. We assume that information about interests of people in different topics is available beforehand. As a simplification, we also assume that it is possible to infer a coarse granularity of broad topics, such that the same topic specific social network can be considered for all events about the topic. For example, a single broad topic for *politics in Pakistan* can be considered for multiple events such as the Emergency, elections, government corruption, etc. Similarly, a broad topic for *machine learning* can be considered for events such as book purchases, notification of new discoveries, interesting applications, etc.

Identification of contextual boundaries: Given an event e, it is assumed that the topic specific social network for the event can be found, and the recipient r can be located in the social network. The task now is to find a contextual boundary for r in this network. Recalling the *strength-of-weak-ties* hypothesis, clusters of people exist in social networks such that members within each cluster have *strong* ties among themselves, and *weak* ties link people across clusters. With respect to the ego of r, weak ties of r are claimed to provide more non-redundant information than strong ties, because the weak ties link r to diverse clusters of people. This is identical to our hypothesis about completeness. Similarly, Hansen's claim that strong ties help understand complex messages much more than weak ties, is identical to our hypothesis about context. Therefore, the goal of the algorithm can be stated to identify clusters of strong ties in topic specific social networks. The contextual boundary of r can now be considered to be the same as the boundary of the cluster to which r belongs. In the same way, the adjacent contextual boundaries can be considered as the adjacent clusters to which links exist through weak ties from r's cluster. Note that considering rigid cluster boundaries is only a simplification for the model; in practice, such contextual boundaries are likely to be more diffuse [32].

Identification of clusters of strong ties in social networks is an actively researched topic. Many algorithms have been developed based on different models for *strong* and *weak* ties. For example, [33] identifies crucial edges in a social network that are responsible to preserve the connectedness of the graph, and then hierarchically identifies clusters which are linked by these edges. [34] identifies locally dense clusters that have a high proportion of cliques or plexes within the clusters. Extensions have also been proposed that consider various edge annotations and weights to denote varying characteristics of ties [35]. Granovetter also proposed a linear combination of four characteristics for tie strength, including the duration of the tie, the intensity of communication, the intimacy, and

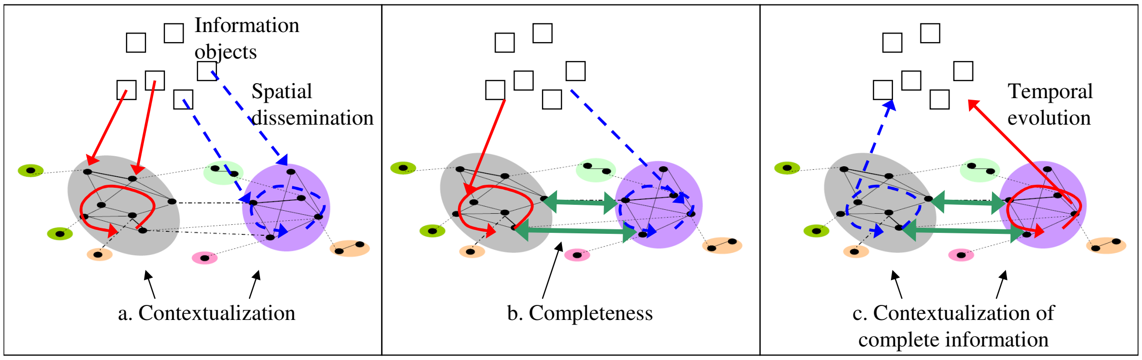

Figure 2.3: Information evolution model

the reciprocity of the tie [27]. We do not propose a new algorithm for cluster identification here. We only experiment with the existing algorithms and analyze which algorithms perform better or worse. More complexity to the algorithms can be added gradually in stages. However, given the clustering output of an algorithm, we define the notion of *strong* and *weak* relationships between people.

Strong relationship: People in the same cluster are categorized as being strongly related to each other even if they do not have a direct connection to each other. It is however assumed that each cluster is a connected graph. Note that a strong relationship exists between any two people in the same cluster, even though they may not have a tie declaration between them.

Weak relationship: People in immediately adjacent clusters which are linked together with weak ties between the clusters, are categorized as being weakly related to each other. Note that a weak relationship exists between any two people in adjacent clusters, even though they may not have a tie declaration between them.

The relationship between all other pairs of people is left as *undefined*. The context and completeness hypotheses can now be restated as follows:

Hypothesis for context restated: For any event e and message recipient r, there exists a set of strongly related people to r, such that related messages about e written by the strongly related people (set of s senders) tend to be more understandable for r than messages written by weakly related people (set of s' senders).

Hypothesis for completeness restated: For any event e and recipient r, the average completeness provided a set of related messages about e written by people strongly related to r (set of s senders), tends to be less than the average completeness provided by messages written by people both weakly and strongly related to r.

2.3.4 Temporal evolution

Based on the definitions given earlier, it is now possible to describe the full model about how and why participatory media would improve media effectiveness. We next define a temporal operator called *contextualization*, which is essential to formulate the model.

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.

The full model is shown in Fig. 2.3 in three steps.

- Part (a) shows that different messages about an event e are read by people in different contextual boundaries of the topic specific social network for that event. We refer to the context considered by the people in the left cluster = c_1, and the context considered by the people in the right cluster = c_2. People within a contextual boundary are strongly related to each other and consider the same context for reading messages: $c_{r1} = c_1$ and $c_{r2} = c_2$. These people 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 people in the same cluster. According to the context hypothesis, this helps people 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 contextual boundaries along the connecting weak ties. This increases the completeness of information about e supplied to people 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 people 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, people in the left cluster may have ignored certain aspects of the event relevant to people in the right cluster $a_2 \backslash a_1$, thinking that those aspects are of no significance to them. However, a contextual comment written by a person in the left cluster about 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, and will form the statement for the third hypothesis of the model.

  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. Therefore, the model is likely to exhibit an emergent behavior with contextual boundaries changing over time, but considering emergent behavior is beyond the scope of this study.

Note that in practice, the temporal evolution can be quite ad-hoc and will not occur in strictly the same time sequence. Therefore, it is hard to test the validity of the exact form in which temporal evolution is modeled, but the model can be tested in terms of its effect on improving media effectiveness:

Hypothesis for temporal evolution: People who participate in blogs and forum discussions are able to develop a better understanding and complete picture about current topics than people who do not use participatory media.

Experiments can be designed to evaluate this hypothesis by using pre- and post-test conditions on different control and test groups of people. It is worth mentioning that a related temporal analysis was done in [36] to trace the changes in political opinion before and after the 1996 presidential elections in USA. The changes were observed with respect to the geographical dispersion of social networks. It was shown that geographically dispersed social ties helped link together different local communities, similar to the weak ties considered in the model. Over time, these weak ties caused significant diffusion of opinion and led to active discussions within local communities, similar to the process of contextualization by strong ties described in the model. This provides considerable support to the validity of the model. Interestingly, it was also observed that as a result of the discussions, the local communities became more divergent in political opinion. This shows that a more complete understanding about current topics may not necessarily lead to consensus, but it can at least lead to more informed formation of opinion.

This model fulfils all the goals stated earlier. It is grounded in previous theory for communication models and social networks. It can be used to precisely define hypotheses that can be tested. It is also flexible enough that complications can be added to improve contextual boundary finding algorithms. A rigorous graph theoretic specification of the model is given in Appendix D.

Chapter 3

Validation methodology

To test the context and completeness hypotheses proposed in the previous chapter, a cohort of message authors and recipients needs to be identified, so that given the nature of relationship between message author and recipient (independent variable), the context and completeness provided by the message for the recipient (dependent variable) can be found. Fig. 3.1 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 relationship between a message author and recipient.
- $\mathbf{U} \in (0, 1)$, denotes the context provided by messages to a message recipient. It is used to infer the distribution of another variable, $(\mathbf{U} \mid \mathbf{R})$, as the context provided by messages to a message recipient, given the nature of relationship between message authors and the message recipient. Referring to the variables defined in the previous chapter, an instance of $\mathbf{U}$ is the average context $E[u]$ provided to a recipient r. For ease of exposition, we use u_i to denote the average context provided to the i^{th} recipient.
- $\mathbf{T} \in (0, 1)$, denotes the completeness provided by messages to a message recipient. It is used to infer the distribution of another variable, $(\mathbf{T} \mid \mathbf{R})$, as the completeness provided by messages to a message recipient, given the nature of relationship between message authors and the message recipient. Referring to the variables defined in the previous chapter, an instance of $\mathbf{T}$ is the average completeness $E[t]$ provided to a recipient r. For ease of exposition, we use t_i to denote the average completeness provided to the i^{th} recipient.

Knowledge of the distribution of $(u_i, R = \{S, W\})$ and $(t_i, R = \{S, W\})$ for all recipients will eventually allow to infer the probabilities $P(U|R)$ and $P(T|R)$. This will indicate the context and completeness provided by the two different types of relationships over the entire population of message recipients. The completeness hypothesis will be validated if $P(T|R = W)$ has more of its mass to the right than $P(T|R = S)$. 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 relationships

[5] Note that although this study falls under the heading of *message effects research*, it is different from most previous studies [37]. We use the nature of relationship between a message author and message recipient to define categories (independent variable) of strong and weak relationships, which produce two kinds of effects (dependent variables) on the message recipients, namely, provide of 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 dependant on the relationship between message authors and recipients.

	Message author-1	Message author-2	Message author-3	....	Message author-n	
Recipient-1	R = S reln,	R = W reln	R = S reln	....	R = S reln	(T=t_1, R=S), (U=u_1, R=S), (T=t_1, R=W), (U=u_1, R=W)
Recipient-2	R = S reln	R = S reln	R = W reln	....	R = W reln	(T=t_2, R=S), (U=u_2, R=S), (T=t_2, R=W), (U=u_2, R=W)
Recipient-3	R = W reln	R = S reln	R = S reln	....	R = W reln	(T=t_3, R=S), (U=u_3, R=S), (T=t_3, R=W), (U=u_3, R=W)
⋮	⋮	⋮	⋮		⋮	⋮
Recipient-m	R = S reln	R = W reln	R = W reln	....	R = S reln	(T=t_m, R=S), (U=u_m, R=S), (T=t_m, R=W), (U=u_m, R=W)
						P(T | R=S), P(U | R=S), P(T | R=W), P(U | R=W)

Figure 3.1: Design framework for study

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 $P(U|R = S)$ has more of its mass skewed to the right than $P(U|R = W)$. 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 contextual boundaries in each topic specific social network. Quantification of u_i and t_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 contextual boundaries within these topic specific social networks.
3. Test the validity of the model using content analysis for the completeness hypothesis, surveys for the context hypothesis, and a mixed-methods approach for the temporal evolution hypothesis.

3.1 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 [38]. 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 contextual boundaries can now be used to label the relationships between the authors and recipients as strong or weak. Eventually, matrices such as the one shown in Fig. 3.1 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. As described in the next chapter, I crawled a large graph of more than 40,000 users from a social networking website. A limitation of automatic social network extraction over manual name generators 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. [39] 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.

3.2 Identification of contextual boundaries

Once the topic specific social networks have been extracted, different clustering algorithms can be used to cluster the graphs [33–35] and label relationships as strong or weak or undefined. Many clustering algorithms require additional tuning parameters to adjust the granularity of clustering. As described in the next chapter, we surveyed a randomly selected sample of users to rate 5 of their randomly selected ties as strong or weak based on communication frequency. We then adjusted the tuning parameters such that the clustering produced by the algorithm was in closest agreement with the ratings given by the users. The hypotheses were then tested with different labeling produced by the algorithms.

3.3 Hypotheses testing

3.3.1 Completeness hypothesis

A single message is taken to be the unit for content analysis. The goal is to determine $(t_i, R = S)$ and $(t_i, R = W)$ 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 [40]. 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.

We feel it is useful to borrow ontology based techniques from computer science that are used in automated semantic analysis of content [41]. An ontology for a topic expresses the relationships between various entities relevant to the topic. A sample ontology for cotton farming in India is shown in Fig. 2.2. It now becomes straightforward to state the rules for content analysis of a messages 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.

Given an ontology, the determination of $(t_i, R = \{S, W\})$ is simple. For a set of messages, t_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 in the previous chapter,

$t_i = \dfrac{|a_M|}{|a_e|.|M|}$. Therefore, the truth of the completeness hypothesis will effectively imply that the fraction of area covered by messages written by authors in the same cluster as the recipient user, is less than the area covered by messages written by authors in the adjacent and same clusters. In other words, the hypothesis states that strong relationships 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 matters as well. Knowledge of $(t_i, R = \{S, W\})$ for different recipients will generate the distribution for $(\mathbf{T} \mid \mathbf{R})$ to test the hypothesis.

Such ontologies will have to be developed for each of the 4 topics being analyzed. I will use public ontology databases such as the Open Directory Project (*www.dmoz.org*) to create ontologies, and also consult experts in each topic for the validity of the ontology. Reliability of content coding will ne ensured by engaging 4 coders for content analysis of the messages along with tests for inter-coder agreement.

3.3.2 Context hypothesis

The context hypothesis is hard to test through content analysis alone. The value of $(u_i, R = \{S, W\})$ 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 [42].

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 the example about the Emergency in Pakistan given earlier:

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. $(u_i, L = S)$ can now be estimated as follows:

$$u_i = \frac{\displaystyle\sum_{j=2}^{5} s_{ij}(\sum_{k=1}^{j-1} s_{ik} + w_{ik})}{\binom{m}{2}}$$

Here, $\binom{m}{2}$ is the total number of pairs of messages. Thus, u_i is the fraction of the number of pairs of messages by strong relationships that promoted more context than other messages. $(u_i, R = W)$ can be calculated similarly, by reversing the values of s and w in the equation above. Knowledge of $(u_i, R = \{S, W\})$ for different recipients will generate the distribution for $(\mathbf{T} \mid \mathbf{R})$ to test the hypothesis. Note that this formulation of u_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 u_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 thumbrules 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 I 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.

Chapter 4

Validation results

The investigation methodology outlined in the previous chapter requires extensive content analysis and surveys, making it unsuitable for us to single handedly validate the model. For this reason, we used a slightly different approach, while preserving the validity and reliability requirements. We crawled a social networking website, Orkut, and designed various graph theoretic measures for context and completeness. We then correlated these measures with short user surveys. The hypotheses for which we tested, have a one-to-one correspondence with the hypotheses described in the previous chapter. We first describe the details of our dataset, and then outline a series of hypotheses to validate the model.

4.1 Preparation of dataset

We selected Orkut for the following reasons.

1. We know from personal experience that Orkut is very popular among the South Asian (India, Pakistan, Sri Lanka) youth population. We use this observation to make an assumption that Orkut gives us access to a reasonably complete real-world social network of people within these cultures.

2. Unlike many other social networking websites, Orkut allows any user to browse any other user's list of friends. This makes it easy to crawl the social network graph.

3. Orkut users can subscribe to various communities of interest and participate in discussions. We assume that membership in a community is evidence of a user's interest in that topic; this allows us to derive topic-specific social networks of users interested in the same topic. The discussions (ie. *messages*) can also be browsed by any user; this allows us to analyze the context and completeness of messages.

4.1.1 Web crawl

We wrote a distributed crawler that could log into the Orkut website by rotating periodically among a number of manually created Orkut usernames. This web-crawler was used to screen-scrape the social network graph of more than 40,000 users. This graph was obtained after pruning a larger

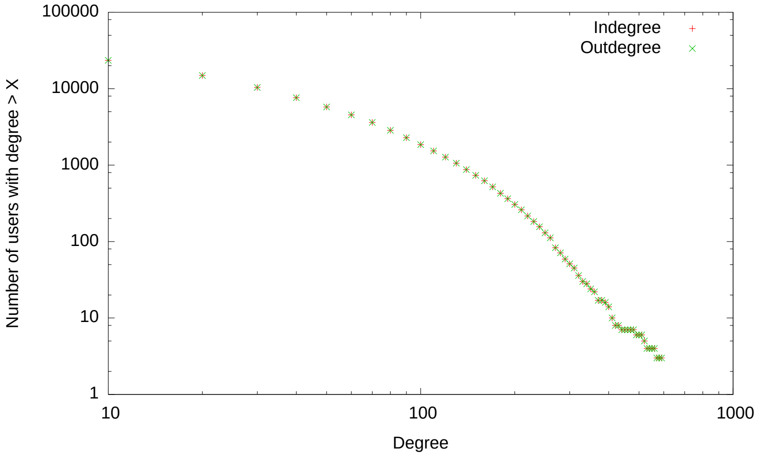

Figure 4.1: Cumulative distribution of user degrees

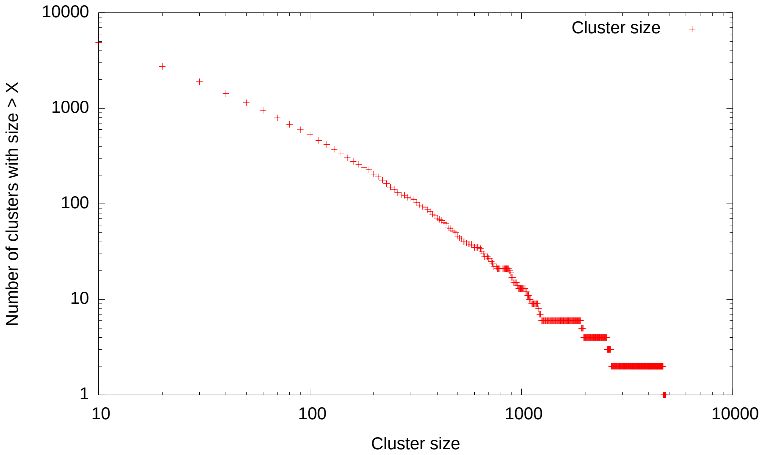

Figure 4.2: Cumulative distribution of topic cluster sizes

graph by retaining a core set of users for whom the difference in indegree and outdegree within the crawled dataset was less than 20. This was done to eliminate dangling edges to users whose social network was not completely known, under the assumption that a real-world social network will mostly have bi-directional links between users [48]. The degree distribution of the core set of users is shown in Fig. 4.1. It was found to have a power-law with a cut-off at 200, similar to observations in other studies [45, 48].

4.1.2 Topic clusters

Orkut users can subscribe to different communities of interest, and the community owners can link their respective communities with other related communities. We crawled this graph of communities in which our core set of users were interested, and then clustered the graph to produce 17,000 interest-based clusters of communities. We refer to these as *topic clusters* henceforth. A stochastic flow simulation graph clustering algorithm [61] was used to produce the clustering. Some examples of topic clusters of communities that were identified are {*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. We found that the distribution of membership-size of users in each topic cluster was a power-law demonstrating the *long-tail* effect among user interests (Fig. 4.2). 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.

4.1.3 Verification methodology

We stored the entire crawled data-set in a MySQL database. To test our hypotheses, we ran network analysis scripts on the data-set to extract graph theoretic measures, and then compared the measured values with data obtained from short user-surveys. A problem we encountered was that Orkut did not allow users to send an arbitrarily large number of messages to other users, and threw up a CAPTCHA [6] to prevent message spam. This posed a significant hurdle for us to send surveys to a large number of users to do a statistically significant study. Therefore, we wrote an application that allowed us to manually enter the CAPCHA phrase whenever required. Using this method, we were able to send almost 6,000 surveys to different users; with a response rate of approximately 20%, this gave us a sufficient number of data-points to test each hypothesis. Although we had to manually tabulate each survey reply, we did not have to reject a lot of data because most replies were comprehensible and given in the format we requested. The response rates also improved with time, probably because many users were repeated and they got more cooperative when they realized that this was a serious study.

4.1.4 Extraction of strong and weak relationships

We assumed that strong and weak relationships can be differentiated from each other based on some clustering algorithm, where links between users in the same cluster are labeled as *strong* links, and links between users in different clusters are labeled as *weak* links. Although significant research exists on the identification of such clusters [53], since we are agnostic to the actual choice of the clustering algorithm, we resort to a simple technique for the purposes of this paper. We use the same stochastic flow simulation graph clustering algorithm [61] used earlier, to cluster the social network graph of all the 40,000 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. 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. We then assumed communication frequency as a proxy for the strength of a tie, and asked the users about the number of times they communicated with their highest and lowest ranked friends in the last three months. Examples of surveys can be founds in Appendix F.1. A total of 314 responses were obtained across the 4 topics, with ratings for 1,473

[6] CAPTCHA is an acronym for *Completely Automated Public Turing test to tell Computers and Humans Apart*. It is commonly used in websites having form-submission pages; an image with distorted alphabets is shown on the screen and the user is asked to type out the alphabets. This is hoped to deter bots from making automated form submissions.

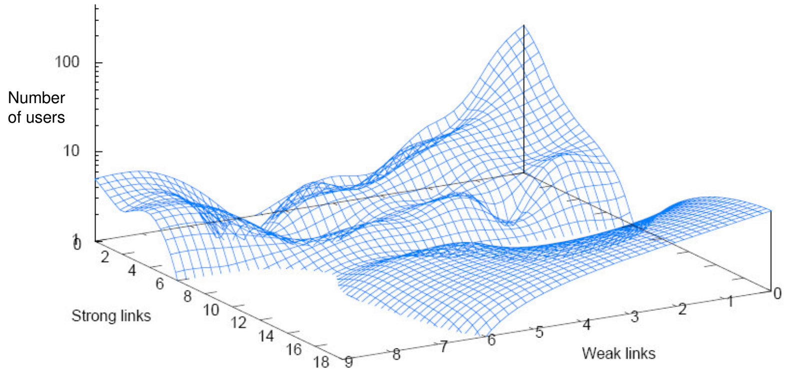

Figure 4.3: Mass-distribution of strong and weak ties in the topic cluster for Orissa

links. The responses were standardized based on communication frequency to obtain interval-based user-ratings for the 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.

Fig. 4.3 shows the mass-distribution of the number of strong and weak links of users over the population of users interested in Orissa. Other topic clusters seem to have very different distribution characteristics, as shown in Appendix F.1.

A limitation of this approach is that a user is allowed to be a member of only one cluster. We are currently designing more general models using probabilistic clustering, which allow users to be members of multiple clusters as well.

4.2 Hypothesis 1: Role of social ties

Strong ties of a user promote context and weak ties promote completeness in the information they receive from their topic-specific social network.

This hypothesis can be considered as an inferior method for validation of the context and completeness hypotheses given in the previous chapter. For example, rather than do a content analysis of messages from different ties of a user, we asked the users for their own opinion about different ties: whether these ties provided them with diverse information, or the ties were familiar with the interests of the user, etc.

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 as follows:

Table 4.1: Comparison of four likely scenarios: {*strong, weak* ties} promote {*context, completeness*}

	Context	**Completeness**
Strong ties	$\mu = .87, n = 133$ $z = -0.48^{***}$	$\mu = .50, n = 133$ $z = -0.51$
Weak ties	$\mu = .35, n = 71$ $z = -4.83$	$\mu = .70, n = 71$ $z = -1.40^{**}$

Context

- 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.

Completeness

- 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.

We received replies from {57, 46, 64, 63} users across the 4 topics, with information about {195, 204, 187, 188} links respectively. We then computed the correlations for four different scenarios, {*strong, weak* ties} promote {*context, completeness*}, and used the z-test to verify which of these scenarios are correct [66]. For each scenario, we performed the z-test by forming the null hypothesis ($\mu = .9$) to indicate that 90% of the subjects believe in the scenario with an error-rate of 10% ($\alpha = 0.1$), and compared it with the alternative hypothesis $\mu < .9$. 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 4.1, and indicate that there is sufficient reason to not reject the two scenarios claimed by the hypothesis: strong ties are indeed more likely to promote context than weak ties, and weak ties are more likely to promote completeness than strong ties. Note that even though the z-test for strong ties promoting completeness is successful, the mean is only 0.5; hence, not considered supportive of the hypothesis. This shows that the *strength-of-weak-ties* hypothesis can be applied to information sharing. Results for experiments with other topics are similar and shown in Appendix F.2.

4.3 Hypothesis 2: Informational quality of a social network

The confidence of a user to rely on members of her topic specific social network to send contextual and complete information to her, can be measured based on the structure of her topic specific social network.

This is meant to give a mathematical form to the notion of context and completeness of social networks of users. Verification of this hypothesis will give further confidence to the validity of the

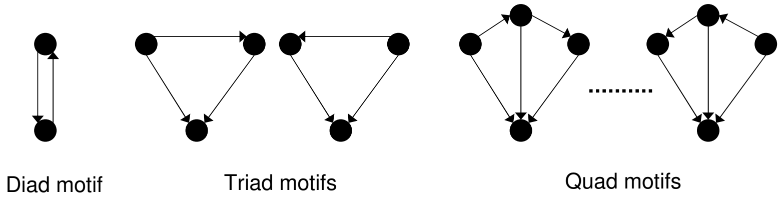

Figure 4.4: Motifs for clustering coefficients

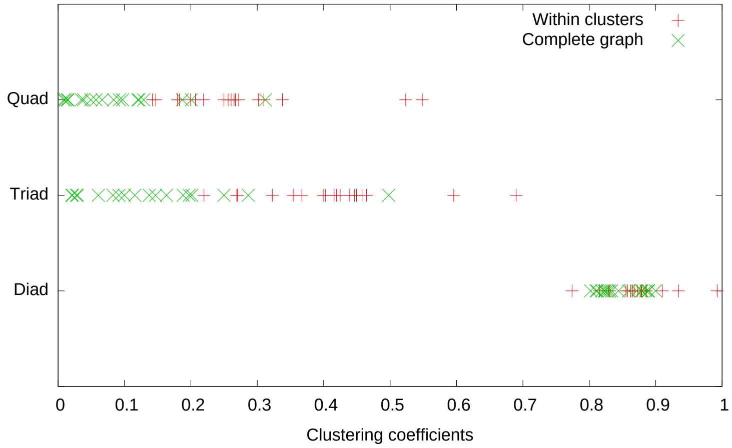

Figure 4.5: Clustering coefficients

model, and indicate that the model can be used to analyze the ability of users to receive contextual and complete information from their social network.

4.3.1 Measurement

1. The social network structure is known in advance as a directed graph $G(U, E)$, where users are represented as nodes U with edges E between users who are friends or know each other.

2. A list of topics T is known in advance, and a boolean value for each (user u_i, topic $t^k \in T$) is also known that indicates whether or not the user is interested in the topic. The social network formed by users who are interested in the same topics is used to form topic-specific social networks.

3. Cluster the topic specific social network such that users within each cluster have strong links between them, and users in different clusters are connected with weak links.

4. For each cluster of strong ties V, calculate its clustering coefficient C_V [58]. We will use the clustering coefficient as a proxy for the ease of flow of information within the cluster. We sometimes refer to C_{V_i} as the clustering coefficient of the cluster of user u_i.

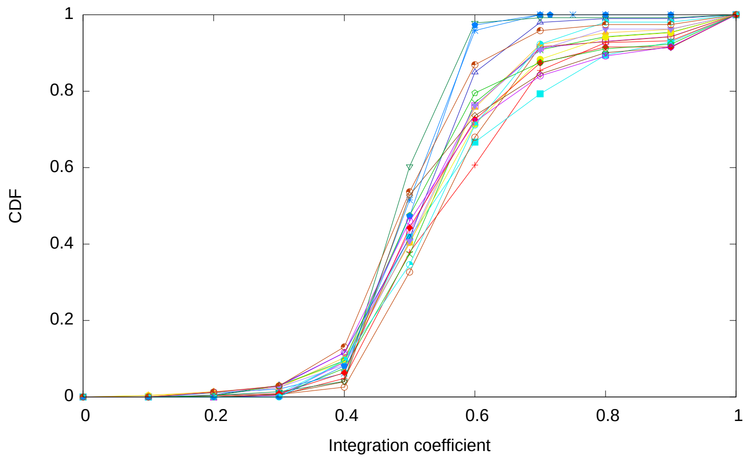

Figure 4.6: Integration coefficients

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

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

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

  Here, d^- is the indegree of person u_i. Thus, define τ_i = maximum number of triangles that can be centered on person u_i.

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

  ie. define $c_i \in [0, 1]$ = clustering coefficient of person u_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 coefficients of all people in this cluster.

The definition of the clustering coefficient proposed in [58] is clearly only one way of quantifying the "density" of a network. Therefore, we experimented with three kinds of motifs, shown in Fig. 4.4. The diad motif indicates the degree of reciprocity between ties. The triad motif indicates the ease of flow of information between a user and two of her ties. Similarly, the quad motif indicates information flow between a user and three of her ties. Fig 4.5 shows the overall clustering coefficients of a few randomly selected topic specific networks, and the mean values of the clustering coefficients 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. However, the triad- and quad-based clustering coefficients are successfully able to discriminate between clusters, as can be seen by the segregation of the coefficients measured within and across clusters in the different topic specific networks. We will use both the triad and quad forms of the clustering coefficient for validation of the hypothesis.

5. For each user, calculate her integration coefficient into her cluster [60]. We will use the integration coefficient of a user as a proxy for the ease of the user to access information from her social network.

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

Here, $d(i,j)$ is the distance from user u_i to u_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 u_i into her cluster V, will be close to 1 for users who are well integrated in their cluster, ie. they are close to many other users. Similarly, γ_i will be close to 0 for users who are present along the boundaries of the cluster and are not well integrated.

Fig. 4.6 shows the cumulative probability distribution of the integration coefficients of the users in the same randomly selected topic specific networks as earlier. The integration coefficient seem to be a suitable metric to reflect the ease of access to information by different users. This is because there is a significant spread in values across different users, and the trend is consistent across different topics as well.

6. Calculate context as follows:

$$Context_i^k = \kappa C_V \ \gamma_i \ |V|$$

Thus, $Context_i^k$ of the social network of user u_i is the product of the clustering coefficient of her cluster, her integration into her cluster, and the size of her cluster. This can be intuitively explained as follows: a high clustering coefficient indicates a high probability that users of this cluster will participate in messages relevant to the topic; a high integration coefficient for a user indicates a high probability that other users will help contextualize information for her; and the larger the size of the cluster, the greater will be amount of contextualization that occurs. κ is a normalization constant to restrict the value of $Context_i^k \in [0,1]$. We will later describe how to estimate κ.

It is an obvious question that should the amount of contextualization be a monotonically increasing function of the cluster size, or should it be represented by a diminishing utility function? Interestingly, it has been observed that group sizes of up to 150 members are sustainable, and it becomes increasing difficult to manage larger groups beyond that [6]. This indicates that context expressed as a linear function of the cluster size is only an approximation for small group sizes, and the extra amount of contextualization derived by a larger group may only lead to a marginal improvement. However, this approximation remains valid in the experiments of the next section because the social network used in the experiments has small sized clusters.

7. Let W_j^i denote weak links from the cluster of user u_i into a neighboring cluster V_j, where V_j is not the same as V_i. Calculate the second-degree integration of user u_i's cluster V_i into cluster V_j, as follows:

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

Here, $d_j(W_j^i, k)$ is the minimum distance to user u_k in cluster V_j from any user $\in W_j^i$. Thus, γ_i^j = the second-degree integration of user u_i's cluster V_i into cluster V_j will be high if the weak links into the neighboring cluster are well distributed across the cluster, such that the minimum distance from a weak tie to every other user in the cluster is small. We will use the second-degree integration coefficient as a proxy for the ease of users in a cluster to access information from an adjacent cluster.

8. Calculate completeness as follows.

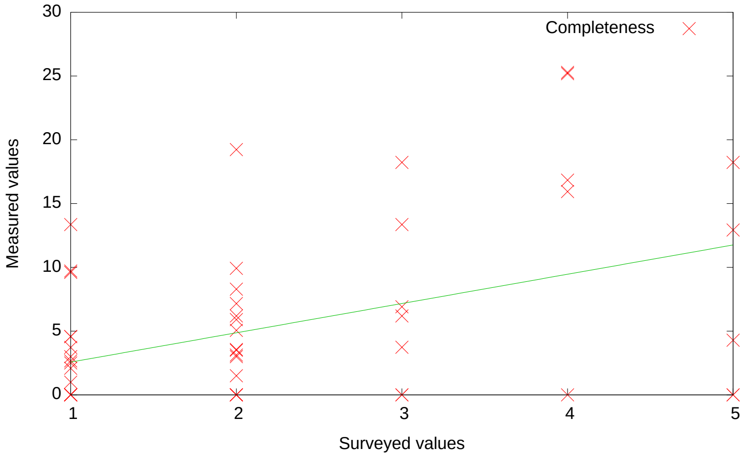

Figure 4.7: Surveyed and measured values of completeness for users interested in Orissa

$$Completeness_i^k = \kappa' \sum_j |V_j| \gamma_i^j$$

Thus, $Completeness_i^k$ is the completeness component of the social network of user u_i, and is expressed as the weighted sum of the sizes of adjacent clusters to which V_i is connected through weak links, the weight being the second-degree integration coefficient of cluster V_i into cluster V_j. This can be intuitively explained as follows: the larger the size of an adjacent cluster, the greater would be the amount of information generated in that cluster; and higher the second-degree integration into that cluster, the greater would be the chances that information from the cluster will flow across. κ' is a normalization constant to restrict the value of $Completeness_i^k \in [0,1]$. We will later describe a method to estimate κ'. Note that we have only included adjacent clusters in the calculation of completeness, although it should be calculated as a feedback-centrality, using fixed-point Eigen value computation. We plan to explore this extended definition in future work.

These measures are time-dependent because the social network structure changes with the addition and removal of links. However, the timescales can be assumed to be fairly large, and we will assume the measures to be static for the purposes of this paper.

4.3.2 Analysis

The same users selected for the previous hypothesis were sent an additional survey; sample surveys are shown in Appendix F.3. The questions were similar to those asked in the previous survey, but were posed directly to the user to assess his/her own abilities to get contextual and complete information from his/her friends. The self-assessed survey values were then compared with the measured values. Although it is not entirely justified to compare ordinal scale ratings with measured values, we do so only as a first step. It is hard to conduct large scale surveys having detailed questions through which we can directly infer the abilities of users to receive contextual and complete information from their social networks.

Fig. 4.7 shows a scatterplot between the measured and surveyed values of completeness for users interested in Orissa. A similar graph is obtained for the triad-based measure of context, shown in

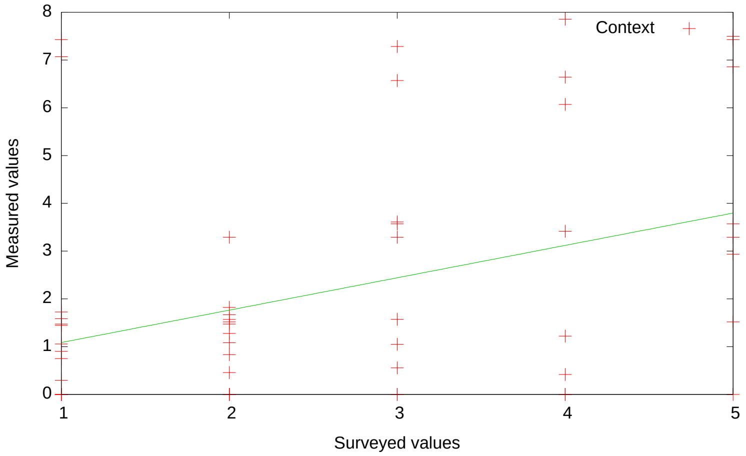

Figure 4.8: Surveyed and measured values of context for people interested in Orissa

Fig. 4.8. Note that the measured values shown in the graph have not been normalized by adjusting κ and κ'. Although the results indicate a positive correlation of 0.50 and 0.53 between the measured and surveyed values of context and completeness respectively, without any normalization it is hard to understand what any arbitrary absolute value of context and completeness actually means. To do this, we used linear regression to estimate the best-fit lines ($y = \beta_0 x + \beta_1$) for both graphs. We then normalized the survey ratings to values between 0 and 1. The measured values were also normalized by dividing them with the y-value of the best-fit line at the survey value of 5 (ie. κ and κ' = respective values of $\beta_0 x + \beta_1$ at $x = 5$). Any values greater than 1, were capped at 1. This made it possible to interpret that values of context or completeness close to 1 are "adequately" high for most users; ie. the unnormalized values can be arbitrarily large depending upon the size of the social network and the number of ties of users, but the highest point of the best-fit lines gives an indication of what are "sufficiently" large values for context and completeness.

Finally, we did a t-test between the normalized values from surveys and measurements, with a null-hypothesis for $\beta_0 = 1$ and the alternative hypothesis for $\beta_0 \neq 1$, with allowable error-rates of $\alpha = 0.1$. According to statistical tables, a t-value between -1.64 and +1.64 is considered as sufficient evidence to not reject the null-hypothesis. The test was successful for both context and completeness for users, with values of 0.43 and -0.54. The same procedure was followed for the quad-based measures of context, and a t-value of -0.002 was obtained. This indicates that the message evolution model can be applied to analyze the abilities of users to receive contextual and complete information from their social network. Results for experiments with other topics are similar and shown in Appendix F.3.

4.4 Hypothesis 3: Context and completeness of messages

Context and completeness of messages can be measured based on the topic specific social network spanned by the message.

Verification of this hypothesis will indicate that the model can be used to analyze the context and completeness of messages. This may seem as an inferior method to validate the context and completeness hypotheses described in the previous chapter, because we do not estimate context and completeness directly through content analysis. Instead, we rely on the opinion of the users to estimate the outcome of these constructs. However, in some sense this hypothesis also indicates a stronger result because we show that we can actually predict the user ratings corresponding to context and completeness.

4.4.1 Measurement

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.

1. Use the topic specific social network obtained earlier to find strong and weak links between users who are part of the *active environment* of message m_r. The active environment refers to the set of users who have written a comment or reply to the message.

2. Calculate context of message m_r for user u_i as follows:

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

Here, the sum is taken over all users u_j in the active environment of the message who are also in u_i's cluster V_i. γ_j is the integration of user u_j into cluster V_i, and C_{V_i} is the clustering coefficient of the cluster of user u_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 user. Context of a message thus indicates the insights contributed by participants with respect to their own strong clusters.

3. Calculate completeness of message m_r for user u_i:

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

The sum is taken over all neighboring clusters V_j, where at least one user from V_j has contributed to the message m_r. As before, 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, it represents the area of the adjacent social network spanned by the message. Assuming that different views are contributed by users in different parts of the social network, completeness of a message thus indicates its diversity of insights.

4. Calculate average context of m_r as follows:

$$Context_{r}^{k} = \kappa \frac{|V_i| C_{V_i} \sum_j \gamma_j}{\sum_i |V_i|}$$

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

5. Calculate average completeness of m_r as follows:

$$Completeness_{r}^{k} = \kappa' \frac{|V_i| \sum_{j} |V_j| \gamma_i^j}{\sum_{i} |V_i|}$$

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

These measures are time-dependent because the active environment of the message changes with time as the message propagates in the social network.

4.4.2 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 users were interested. This data was suitable for testing the previous two hypotheses 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.

A problem we encountered was that Orkut did not list more than 1,000 community members at any one time. Therefore, we ran 3 crawls for each community at different times to find enough users to test the hypothesis, and aggregated the lists of members with an additional crawl of all users participating in the selected messages. This is sufficient for testing the hypothesis because our formulations for context and completeness of messages depend only upon the network spanned by the active environment of the message.

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 (ie. message context), and if the discussion covered other diverse circumstances for use of the law (ie. message completeness). A 5-point scale was used for the ratings. We sent a survey for each message to 15 users, and received 837 replies. The ratings given in the replies were then compared with the measured values of context and completeness of the message for the users. Although this method suffers from the same criticism as earlier of using an ordinal scale as an interval scale, we feel this is suitable as a first step.

We now calculated the normalization constants κ and κ' by finding the best-fit line between the measured and surveyed values for different users, similar to how it was done for the previous hypothesis. The same constants were then used to normalize the measured values for average context and completeness in the message. Any values greater than 1 after the normalization, were capped at 1. Fig. 4.9 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 then 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

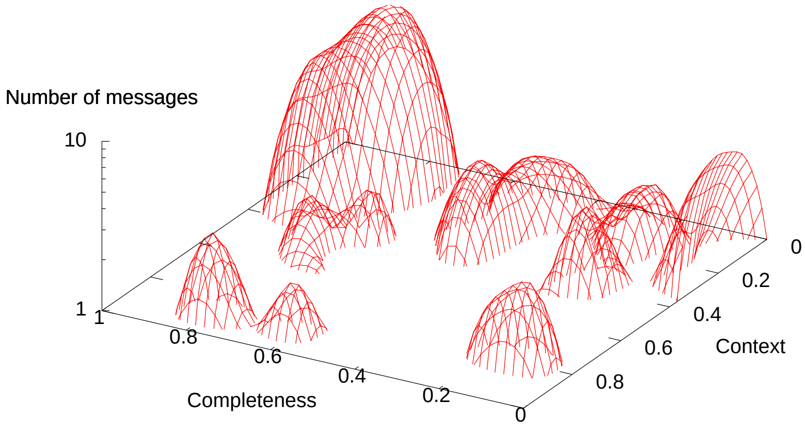

Figure 4.9: Mass-distribution of triad-based context and completeness of all messages

Table 4.2: 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	$\rho_n = .58$ $\rho_m = .90$	$t_n = -3.01^{*}$ $t_m = 1.09^{***}$

of the context and completeness ratings given by the users. The same procedure was followed for quad-based measures of context.

The correlation coefficients and t-test values are shown in Table 4.2 and Table 4.3 (ρ_n and t_n for context, ρ_m and t_m for completeness). 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. This gives further credibility to the model, and shows that the model can be used to analyze the context and completeness of messages.

4.5 Hypothesis 4: Temporal evolution of messages

Changes with time in context and completeness of messages can be perceived by users, and correlated with measurements based on the topic-specific social network spanned by the message.

Table 4.3: 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	$\rho_n = .68$ $\rho_m = .88$	$t_n = -0.75^{***}$ $t_m = 0.04^{***}$

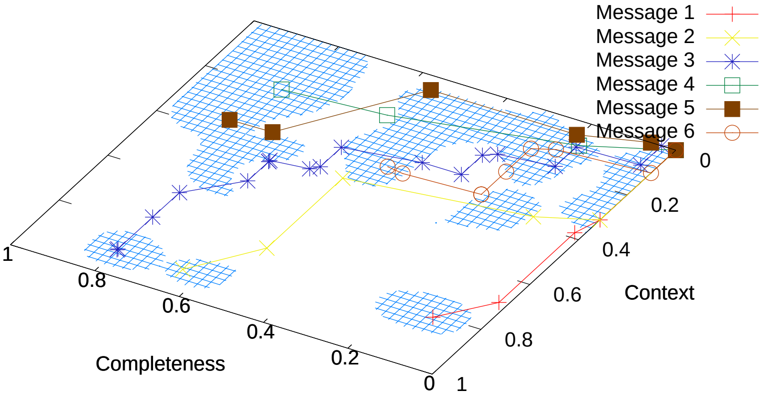

Figure 4.10: Evolution of context and completeness of messages

Table 4.4: Surveyed and measured values for changes in context and completeness of messages

	Context	**Completeness**
Large change	$\mu = .92, n = 19$ $z = 0.17^{***}$	$\mu = .87, n = 23$ $z = -0.22^{***}$
No change	$\mu = .75, n = 32$ $z = -0.98^{***}$	$\mu = .53, n = 35$ $z = -2.20^{*}$

Fig. 4.10 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 from Fig. 4.9 projected onto the XY plane. Each message now 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.

From across all messages, we randomly selected 94 messages, and 2 replies in each message, where the replies caused the context or completeness of a message to change by more than 0.1 (ie. a large change) or less than 0.05 (ie. almost no change). We then sent out a survey specific to each message to 10 users from the same communities, and asked them to rate the replies in terms of whether or not they promoted context or completeness. The questions were framed in the same way as for the previous hypothesis.

Out of 304 replies we received, we took the weighed mean of the ratings for each reply and labeled them as causing {*no change, large change*} in the context and completeness of messages. We then correlated this with the values inferred through measurements for *no change* and *large change*, and tested for statistical significance using the z-test. The results are shown in Table 4.4. Although the tests do not succeed in a few scenarios, they do indicate that it is possible to track changes in the context and completeness of messages by observing changes in the active environment of the message. This lends further evidence for the correctness of the model.

4.6 Limitations of evaluation

Our current approach for testing the hypotheses through large scale user-surveys suffers from certain limitations, chiefly due to the brevity of surveys. First, a user's rating for other users are typically influenced by many external factors such as the amount of credibility of the users, their geographical proximity, etc. Longer surveys can be designed to factor out the influence of such external parameters, but we have ignored it for simplicity. The second shortcoming is the correctness of using ordinal ratings for z-tests and t-tests without converting them to an interval scale. This problem can again be addressed by longer knowledge-based surveys that try to directly assess the context and completeness of messages and social networks of users, but it would come at the cost of increased complexity and lower response rates from users. Third, it is seen that the hypotheses tests are successful in most but not all the cases. We again feel that this is because of the brevity of surveys; our normalization techniques could have introduced significant noise into the results. Fourth, we recognize that our evaluation may not be generalizable to other kinds of social networks. However, the propositions made in this study clearly have much broader implications, and further analysis of the theoretical constructs of context and completeness will lead to better generalizability.

Chapter 5

Applicability of the model

The model and related theoretical constructs of context and completeness discussed in the previous chapters have many applications. We next discuss a few extensions to the model, and build upon the insights to propose some useful applications.

5.1 Information usefulness

Context and completeness can be considered as components for the usefulness of information, that is, a person may find a particular message to be useful either because it provides more completeness to a recipient, or it provides more context, or both. This insight is in agreement with other research in the field of information science. Information scientists have explored the notion of useful information in a message, and characterized it through features such as comprehensibility of the message, its scope, freshness, accuracy, credibility, and topic [7, 8]. Similarly, media researchers have explored the "effects" of information, and use terms such as resonance, simplification, repetition, and opinion diversity to describe the features of a message that can explain the effects produced by the message [9]. Context and completeness have a one-to-one correspondence with these features.

Context: This relates to the ease of understanding of the message, based on how well the message content explains the relationship of the message to its recipient. *Comprehensibility* of the message [7], or *simplification* of the meaning of the message [9], can be considered as outcomes of the amount of *context* in the message. That is, messages that are more contextual for users, will be more comprehensible and simple for them.

Completeness: This denotes the depth and breadth of topics covered in the message. A concrete definition of *depth* and *breadth* is proposed in [52], as the depth and breadth of the topic ontology graph covered by the message. The *scope* of the message [7], or the *opinion diversity* expressed in the message [9], can be considered as outcomes of the amount of *completeness* in the message.

Thus, messages having a greater amount of context and completeness can be considered to be more useful for a message recipient. However, context and completeness of messages cannot be derived in a straightforward manner through semantic content analysis alone. We have shown in this study that these features are influenced by user participation, and can be measured based on the social network formed by the participants and readers when the message *evolves* with time. Different message recipients are likely to attach different priorities to each of these features, and this notion can be used to enhance existing ranking metrics for messages [44]. We explored this idea by developing a personalized Bayesian usermodel for each user to predict the order and extent of recommending information about some topic to the user [64].

Apart from context and completeness, the topic-relevance of messages, and credibility and expertise of message authors are other important features to determine message usefulness. We have already taken topic relevance into account by considering only topic specific social networks. We next briefly discuss the role of credibility in information usefulness.

Credibility: Message or user credibility can be considered as the amount of trust in the message or message source respectively, as perceived by the message recipient. Models for the calculation of related concepts such as trust and reputation have been proposed in [50, 55, 56], and can be adapted to this scenario. However, it is important to distinguish between the credibility of a message and the reputation of its author. An author gains reputation by producing credible information, and the credibility of information is calculated based on ratings given by various users. Convergence can be ensured by propagating credibility and reputation scores on the hyperlinks and trackbacks between various message such as blogs, videos, etc. Note that credibility itself is contextual because, considering the news article example given in the first chapter, people in group A may not find the second comment to be credible, whereas people in group B may consider it otherwise. For this reason, social network information should be incorporated in credibility computation to derive different kinds of credibility scores based on the local social network neighborhood, the global neighborhood, and history of the information itself [65]. Credibility values can then be accommodated as weighting factors in our formulae of context and completeness for messages and the social network of users.

5.2 Ranking metrics for messages

A preliminary verification of the intuition that the usefulness of a message will increase with its context and completeness, can be done by defining a ranking metric for messages as follows:

1. Context + μ Completeness: $\mu \in (0, 1)$

2. Context

3. Completeness

4. log(Context) + μ log(Completeness)

Unfortunately, it is difficult to test ranking metrics for messages because it is hard to find a sufficient number of similar messages so that their ranks are not influenced by other factors such as the message topic or freshness of the discussion. For this reason, instead of ranking messages, we used the same metrics to rank communities by calculating the average context and completeness values of the messages in the communities. We then selected 5 communities in each of the 4 topics, calculated the average context and completeness of messages in each community, and compared the rankings based on our metrics with rankings given by user-surveys. The user rankings were obtained by asking 200 randomly selected users to rank the 5 communities in different topics by examining the discussions in each community.

{10, 15, 11, 12} sets of rankings were obtained for communities in the 4 topics respectively. The measured rankings were then compared with the user-rankings using the Kendall and Spearman rank correlation tests [59]. Other ranking schemes such as the number of users in the community,

Table 5.1: Comparison of ranking schemes for communities

	Economics	**Orissa**	**Books**	**Mumbai**
Number of members	$\rho = .09, \tau = .11$ $-z_\rho = 8.69, -z_\tau = 10.1$	$\rho = .73, \tau = .66$ $-z_\rho = 0.96, -z_\tau = 1.44$	$\rho = .72, \tau = .63$ $-z_\rho = 1.04, -z_\tau = 1.76$	$\rho = -0.01, \tau = -0.07$ $-z_\rho = 15.1, -z_\tau = 21.2$
Freq of discussions	$\rho = -0.52, \tau = -0.27$ $-z_\rho = 10.2, -z_\tau = 14.1$	$\rho = -0.15, \tau = -0.04$ $-z_\rho = 14.4, -z_\tau = 11.1$	$\rho = .21, \tau = .45$ $-z_\rho = 8.01, -z_\tau = 3.89$	$\rho = .12, \tau = .47$ $-z_\rho = 17.8, -z_\tau = 6.35$
Context	$\rho = .14, \tau = .07$ $-z_\rho = 7.20, -z_\tau = 9.01$	$\rho = .51, \tau = .44$ $-z_\rho = 2.79, -z_\tau = 3.84$	$\rho = .78, \tau = .69$ $-z_\rho = 0.67, -z_\tau = 1.32$	$\rho = .02, \tau = -0.05$ $-z_\rho = 18.9, -z_\tau = 31.8$
Completeness	$\rho = .68, \tau = .59$ $-z_\rho = 1.34, -z_\tau = 2.09$	$\rho = .59, \tau = .44$ $-z_\rho = 2.08, -z_\tau = 4.14$	$\rho = .69, \tau = .55$ $-z_\rho = 1.32, -z_\tau = 2.80$	$\rho = .87, \tau = .87$ $-z_\rho = 0.14, -z_\tau = 0.15$
Context + 2 . Completeness	$\rho = .73, \tau = .71$ $-z_\rho = 0.89, -z_\tau = 1.02$	$\rho = .67, \tau = .55$ $-z_\rho = 1.38, -z_\tau = 2.38$	$\rho = .69, \tau = .56$ $-z_\rho = 1.32, -z_\tau = 2.81$	$\rho = .79, \tau = .69$ $-z_\rho = 0.62, -z_\tau = 1.29$
Context + 0.5 . Completeness	$\rho = .63, \tau = .57$ $-z_\rho = 1.52, -z_\tau = 2.08$	$\rho = 0.66, \tau = 0.56$ $-z_\rho = 1.46, -z_\tau = 2.36$	$\rho = .78, \tau = .69$ $-z_\rho = 0.67, -z_\tau = 1.32$	$\rho = .79, \tau = .69$ $-z_\rho = 0.62, -z_\tau = 1.29$
log Context + log Completeness	$\rho = .63, \tau = .57$ $-z_\rho = 1.52, -z_\tau = 2.08$	$\rho = .66, \tau = .56$ $-z_\rho = 1.46, -z_\tau = 2.36$	$\rho = .78, \tau = .69$ $-z_\rho = 0.67, -z_\tau = 1.32$	$\rho = .79, \tau = .69$ $-z_\rho = 0.62, -z_\tau = 1.39$

and the frequency of posting new discussions, were also compared. A z-test was then done with the mean of the correlation coefficients for different user rankings. The results are shown in Table 4.5. Metrics based on context and completeness consistently give better correlations than other metrics. The following metrics produced the best results for the 4 topics:

1. *Economics*: (Context + 2 . Completeness) gave the best result. In general, completeness seems to have a greater effect on the ranking.

2. *Orissa*: (Number of members) produced the best result, but metrics based on context and completeness were also close and similar to each other.

3. *Books*: (Context), (Context + 0.5 . Completeness), and (log Context + log Completeness) gave the same results. Context seems to have a greater effect on the ranking.

4. *Mumbai*: (Completeness) gave the best result, and seems to have a greater effect on the ranking.

This shows that the insights gained from the model can be used to develop ranking metrics for usefulness of messages. It is also interesting to note that different values of μ are good for different topics, indicating that the mix of context and completeness desired by users is likely to be topic-dependent.

5.3 Information capital

We find our notion of context and completeness to be very similar to that of *social capital* in sociology. *Social capital* is defined as a network characteristic emerging from the *resources* or *social assets* available with people in a social network, that are made available to other people or communities in the social network [22, 24]. For example, these resources can include factors such as the level of education of members in a community, or their economic status, or their helpful nature, etc. This is described in more detail in Appendix E. The overall ability of the community to gain access to these resources depends upon the structure of the social network. If information is considered as a resource available with different people, then the ability of people to access this information can be considered similar to social capital. We term this as *information capital*, and interestingly, this is exactly what we have measured in Section 4.3 through our notions of context and completeness of the social network of users!

We propose that information capital should be written as a tuple of (*context*, *completeness*), such that application scenarios can specify the relative importance of context and completeness. For example, information capital related to personal communication is likely to be dependent only upon context, whereas information capital related to objective factual information may be dependent only upon completeness, and information capital related to subjective information may be dependent upon both context and completeness. Recall that the topic-specific dependence of the relative importance of context and completeness was also noticed in Section 5.2, where different values of μ resulted in better correlations for different topics. This insight poses an interesting research question to characterize the relative amounts of context and completeness that are optimal for information sharing in different scenarios.

The same ideas can be extended to *information goodness*, which indicates the amount of contextual and complete information that is provided by the message: messages having more context and completeness make more more "good" information available to people. Some concepts related to information capital and information goodness are described next.

Group information capital: Information capital can easily be generalized from an individual notion to a group notion. This will indicate the ability of groups instead of individuals, to share contextual and complete information among themselves.

Non-conservative nature: Information capital and goodness are not conserved. This means that a person's information capital does not diminish by sharing her information with others. In fact, the more that people share information, more will be their information capital. Similarly, information goodness increases as more and more people contribute to a message. Information capital and goodness thus appear to be non-conserved unbounded measures, but bounds do exist due to the limited processing capabilities of people, such as the maximum number of friends a person can have or the maximum amount of information she can consume in a day. This poses an interesting challenge to further develop the concepts of information capital and goodness by taking such psychological and physiological factors into account.

Relationship between information capital and information goodness: Although information capital has been defined so far as the ability of people to receive contextual and complete information being shared in their social network, it does not imply that the people will make use of the information that is made available to them, ie. view it or read it. This can be understood by differentiating between *normative* and *descriptive* information capital. The terms are borrowed from literature on decision theory, where *normative* signifies information capital as defined so far, ie. "what it ought to be", and *descriptive* signifies information capital as "what it actually is", ie. based on the observed user-behavior.

Descriptive information capital of a person can be considered as the sum of the goodness of information received by the person. Thus, if $Context_{ir}^{k}$ and $Completeness_{ir}^{k}$ denote the personalized context and completeness for message m_r of topic t^k for user u_i, as calculated earlier, then the descriptive information capital of u_i can be calculated as:

$$\sum_{m_r \in L_i^k} (\text{Context, Completeness})\ l_r$$

Here, L_i^k is a list of messages of topic t^k received in the past by person u_i, and l_r is a measure of how useful the person found message m_r to be. We assume that it is possible to approximate

l_r based on ratings given by the person, or the amount of time spent on the message, etc. This definition continuously increases with time and does not take the freshness of information into account; therefore, moving average based measures can be developed that discount the information goodness obtained from old messages.

Ideally, normative and descriptive information capital should converge to the same values. This is likely never to be true; descriptive information capital will most often be an under-estimate of the normative information capital. But the difference can be indicative of what "is" and what it "ought" to be, ie. how much "good" information does a person actually receive, versus how much ability does the person have to receive "good" information.

5.4 Information retrieval applications

We mention a few applications we plan to build in the future, that can benefit from the insights developed in this paper.

Recommender systems: Existing recommender systems for movies, books, blogs, etc seem to only focus on credibility of the reviews to rank them, and ignore the role of context and completeness of reviews. Knowledge of the social network connecting reviewers and customers can help estimate the context and completeness of reviews, and improve the overall usefulness of the review page.

Growing a social network: The notion of topic-specific information capital can be used to suggest new links between users, which if formed can increase their information capital, that is, their ability to receive contextual and complete information from their social network. For example, on professional online social networks such as LinkedIn, users having high information capital for a particular topic but low information capital for a different topic, could benefit from linking to user who have complementary interests.

The model presents a useful categorization of the abstracts concepts of context and completeness, and can be used to conceive many more applications.

5.5 Related work

To the best of our knowledge, we are not aware of any prior research that has examined similar evolutionary characteristics of context and completeness for participatory messages. We therefore attempt to situate our work in 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, [48] studied the link structure of users of four online social networking websites. Similarly, [50] studied the link structure of internal blogs of a large corporation. [51] 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, [47] used social networks to improve web-page rankings produced by Google.Com. [49] 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, [45] 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. [46] 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.

Chapter 6

Conclusions

In this study, we proposed a social network based model of how people perceive and interpret participatory media content. The model explains the underlying processes of online public discourse through which the effectiveness of news media is enhanced to gain a better understanding of topics discussed in mass media, and to present people with diverse viewpoints to avoid media bias. We validated the model through a series of hypotheses using measurements and user surveys of an online social networking website. Two key theoretical constructs of context and completeness of information were introduced, which play a role in indicating the usefulness of information for a person. Graph theoretic measures for these constructs were then developed and validated. The mathematical form of the theory makes it usable to improve the design of information search and recommendation systems on the Internet.

Bibliography

[1] M. Brown, “Abandoning the News,” Carnegie Reporter, Vol. 3, No. 2, 2005.

[2] D. Sifry, “The State of the Live Web,” http://www.sifry.com/alerts/archives/000493.html, April 2007.

[3] “Blog Readership,” http://www.stateofthemedia.org/, 2007.

[4] Dan Gillmor, “We the Media,” O’Reilly Media, Sebastopol, USA, 2004.

[5] Howard Rheingold, “Smartmobs,” Basic Books, Cambridge, USA, 2002.

[6] Malcolm Gladwell, “The Tipping Point,” 2000, Little, Brown, and Company, New York, USA.

[7] K. Maglaughlin and D. Sonnenwald, “User Perspectives on Relevance Criteria: A Comparison among Relevant, Partially Relevant, and Not-Relevant Judgements,” Journal of the American Society for Information Science and Technology, Vol. 53, No. 5, 2002.

[8] 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.

[9] J. Bryant and D. Zillman, “Media Effects: Advances in Theory and Research,” Lawrence Erlbaum Associates, New Jersey, USA, 2002.

[10] “Blog Action Day,” http://blogactionday.org/, 2007.

[11] Edward Herman and Noam Chomsky, “Manufacturing Consent,” Pantheon Books, New York, USA, 1988.

[12] 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.

[13] 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.

[14] Jurgen Habermas, “The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society,” MIT Press, USA, 1989.

[15] L. Sjoberg, “You and Your LiveJournal and You,” http://www.wired.com/gadgets/miscellaneous/commentary/alttext/2006/06/71142, 2006.

[16] G. Piccalo, “BoingBoing.Net Bounced into TV Territory,” http://www.latimes.com/entertainment/news/newmedia/la-et-boing3oct03,0,2256164.story, 2007.

[17] Maeve Cooke, “Language and Reason : A Study of Habermas’s Pragmatics,” MIT Press, USA, 1994.

[18] A. de Moor and L. Efimova, “An Argumentative Analysis of Weblog Conversations,” Proc. Language-Action Perspective on Communication Modelling, 2004.

[19] G. Gerbner, “Toward a General Model of Communication,” Proc. Audio-Visual Communication Review, 1956.

[20] Denis McQuail, “Communciation Models for the Study of Mass Communication,” Longman Group, UK, 1981.

[21] Marshall McLuhan, “Understanding Media,” Gingko Press, Corte Madera, USA, 1964.

[22] N. Lin, “Building a Network Theory of Social Capital,” International Sunbelt Social Network Conference, 1999.

[23] R. Putnam, “Bowling Alone: The Collapse and Revival of American Community,” Simon and Schuster, New York, USA, 2000.

[24] A. Krishna, “Active Social Capital: Tracing the Roots of Development and Democracy,” Columbia University Press, New York, USA, 2002.

[25] D. Narayan, “Bonds and Bridges: Social Capital and Poverty,” Poverty Group, World Bank, 1999.

[26] N. Ellison, C. Steinfield, and C. Lampe, “Spatially Bounded Online Social Networks and Social Capital: The Role of Facebook,” Proc. International Communication Association, 2006.

[27] M. Granovetter, “The Strength of Weak Ties,” American Journal of Sociology, Vol. 78, No. 6, 1973.

[28] M. Granovetter, “The Strength of Weak Ties: A Network Theory Revisited,” Sociology Theory, Vol. 1, 1983.

[29] M. McPherson, L. Smith-Lovin, and J. Cook, “Birds of a Feather: Homophily in Social Networks,” Annual Review of Sociology, Vol. 27, 2001.

[30] M. T. Hansen, “The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits,” Administrative Science Quarterly, Vol. 44, 1999.

[31] P. Sainath, “Give us a Price, not a Package,” http://www.indiatogether.org/2006/aug/psa-price.htm, Aug 2006.

[32] H. Eulau and L. Rothenberg, “Life Space and Social Networks as Political Contexts,” Political Behavior, Vol. 8, 1986.

[33] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” Proc. National Academy of Sciences, Vol. 99. No. 12, 2002.

[34] S. Wasserman and K. Faust, “Social Network Analysis,” Cambridge University Press, UK, 1994.

[35] 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.

[36] B. Baybeck and R. Huckfeldt, “Urban Contexts, Spatially Dispersed Networks, and the Diffusion of Political Information,” Political Geography, Vol. 21, 2002.

[37] Sally Jackson, “Message Effects Research: Principles of Design and Analysis,” Guilford Press, New York, USA, 1992.

[38] B. Hogan, J. Carrasco, and B. Wellman, “Visualizing Personal Networks: Working with Participant-Aided Sociograms,” Field Methods, Vol. 19, No, 2, 2007.

[39] 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.

[40] Daniel Riffe, Stephen Lacy, and Frederick Fico, “Analyzing Media Messages: Using Quantitative Content Analysis in Research,” Lawrence Erlbaum Associates, New Jersey, USA, 2005.

[41] T. Gruber, “Ontology,” To appear in *Encyclopedia of Database Systems*, L. Liu and M. Tamer zsu, 2008, Springer Verlag, USA.

[42] 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.

[43] John Creswell, “Research Design: Qualitative, Quantitative, and Mixed Methods Approaches,” Sage Publications, Thousand Oaks, USA, 2003.

[44] C. Manning, P. Raghavan, and H. Schtze, “Introduction to Information Retrieval,” 2008, Cambridge University Press, USA.

[45] L. Adamic and E. Adar, “How to Search a Social Network,” Social Networks, Vol. 27, No. 3, 2005.

[46] J. Kleinberg, “The Small-World Phenomenon: An Algorithmic Perspective,” In Proc. STOC, 2000.

[47] A. Mislove, K. Gummadi, and P. Druschel, “Exploiting Social Networks for Internet Search,” Proc. Hotnets, 2006.

[48] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and S. Bhattacharjee, “Measurement and Analysis of Online Social Networks,” Proc. IMC, 2007.

[49] X. Song, B. Tseng, C. Lin, and M. Sun, “Personalized Recommendation Driven by Information Flow,” Proc. SIGIR, 2006.

[50] P. Kolari, T. Finin, K. Lyons, Y. Yesha, Y. Yesha, S. Perelgut, and J. Hawkins, “On the Structure, Properties, and Utility of Internal Corporate Blogs,” Proc. ICWSM, 2007.

[51] 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.

[52] X. Zhu and S. Gauch, “Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web,” Proc. SIGIR, 2000.

[53] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe, “A Framework for Community Identification in Dynamic Social Networks,” Proc. SIGKDD, 2007.

[54] D. Kempe, J. Kleinberg, E. Tardos, "Maximizing the Spread of Influence Through a Social Network," Proc. SIGKDD, 2003.

[55] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of Trust and Distrust," Proc. WWW, 2004.

[56] J. M. Pujol, R. Sanguesa, and J. Delgado, "Extracting Reputation in Multi Agent Systems by Means of Social Network Topology," Proc. AAMAS, 2002.

[57] A. Das, M. Datar, A. Garg, and S. Rajaram, "Google News Personalization: Scalable Online Collaborative Filtering," Proc. WWW, 2007.

[58] M. E. J. Newman, "The Structure and Function of Complex Networks," SIAM Review, Vol. 45, No. 2, 2003.

[59] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top-k Lists," Proc. ACM-SIAM Symp. on Discrete Algorithms, 2003.

[60] T. Valente, "Network Models for the Diffusion of Innovations," 1995, Cresskill Hampton Press, USA.

[61] S. Dongen, "MCL: A Cluster Algorithm for Graphs," PhD thesis, University of Utrecht, 2000.

[62] D. Watts, P. Dodds, and M. Newman, "Identity and Search in Social Networks," Science, Vol. 296, No. 5571, 2002.

[63] A. Seth, "An Infrastructure for Participatory Media," Proc. AAAI Workshop on Recommender Systems, 2007.

[64] A. Seth and J. Zhang, "A Social Network Based Approach to Personalized Recommendation of Participatory Media Content," Manuscript under review, 2007.

[65] B. Fogg and H. Tseng, "The Elements of Computer Credibility," Proc. SIGCHI, 1999.

[66] L. Wasserman, "All of Statistics: A Concise Course in Statistical Inference," 2004, Springer-Verlag, New York, USA.

Appendix A

Public and private spheres

Habermas suggests that traditionally the private sphere of families and friends interacted through personal letters, *salon* discussions, and home gatherings, to create public opinion and promote a humanitarian perspective in public services provided by the governments in social welfare states. However, the creation of the modern bourgeoise society transformed the spheres when many private bourgeoise individuals (such as the capitalists) gained representation in governments, and started influencing public services for their personal gain. Around the same time, the decentralized *salon* discussions lost their importance to the centralized printing press as the modern agency legitimized to represent public opinion. But centralization of the press made it easier for the bourgeoise to "privatize" it to represent their own interests. Furthermore, education became a prerequisite for private individuals to gain membership in the bourgeoise society to represent their interests.

Therefore, for the first time, formation of public opinion moved out of the traditional *salons* that had been centers for interaction among the private individuals. The bourgeoise now referred to themselves as the new "private" sphere, and the traditional private sphere lost its representation in public opinion. News was essentially converted into a commodity that shaped public opinion for the interests of the bourgeoise individuals. The subsequent creation of civil society organizations tried to gain back representation for the traditional private sphere through strategic partnerships, but this further caused news to be commoditized by essentially trading news with materialistic gains. This changed the entire meaning of public opinion because it did not represent the "public" any more. Even after the press evolved from print to electronic forms of mass media, public opinion remained as a commodity.

Appendix B

Communicative action

Habermas uses the two-level Marxian model to differentiate between the *lifeworld* for social reproduction of values and culture, and the *system* for materialistic reproduction of economic goods and services. All communication between agents in each level is considered to be aimed at *rationalizing* the objectives of the agents in that level. Communication in the system is through *instrumental action*, aimed at the strategic positioning of arguments by agents to exploit the other agents for materialistic gains. Communication in the lifeworld is through *communicative action*, aimed at reaching a common understanding about some topic. Habermas claims that communicative action between agents can be rationalized by following a set of discourse rules for arguments put forth by the agents. For example, agents are allowed to freely challenge the arguments given by other agents, who will then have to justify the arguments they put forth. Discourses obeying these rules will eventually lead to a common understanding among the agents.

Appendix C

General model of communication

Fig. C.1 shows Gerbner's general model of communication [6]. A real event **E** is perceived by an observer **M** as event **E**′, depending upon the individual perceptions of **M**. **M** now generates a message **SE** about the event. Here, **S** denotes the form of the content (for example, a picture, or spoken words, or a written article), and **E** denotes the actual content expressing information about the event. This model can now be treated as a building block to form a chain of communication. Thus, another observer $\mathbf{M}^2$ may see or hear or read **SE** depending upon its form, interpret it as **SE**′, produce another message **SSE**, and so on. In fact, feedback to the original observer can also be modeled by feeding **SSE** back to the observer as an event.

Gerbner's model is flexible enough to capture many aspects of communication including perceptual aspects of the observer depending upon his context and selection bias, temporal aspects of information flow and feedback, and the type of the communication channel. Earlier models by Shannon and Lasswell did not include feedback or perceptual aspects of communication [20]. Schramm's model included perceptions and feedback, but it was not flexible to generalize to multiple links of information flow [20].

[6] Figure obtained from http://www.cultsock.ndirect.co.uk/MUHome/cshtml/introductory/gerbner.html

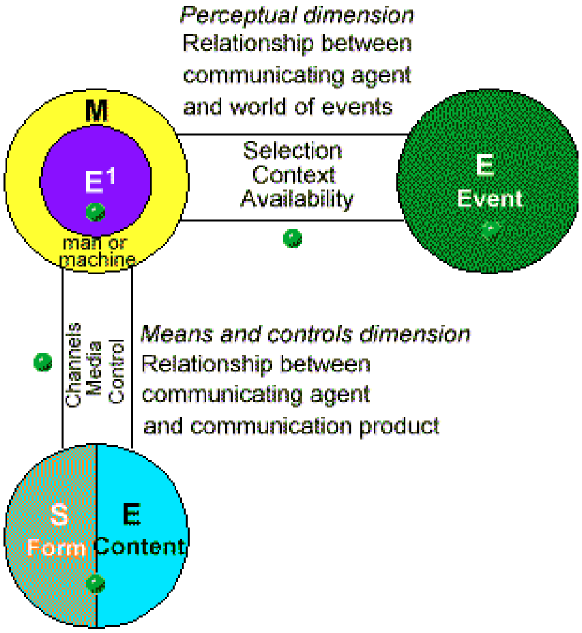

Figure C.1: Gerbner's General Model of Communication

Appendix D

Graph theoretic model specification

Define the following matrices:

- **P**: $p_{ij} = 1$ if person p_i is linked to person p_j, 0 otherwise. This matrix can be obtained through name generators, or social networking websites. The definition can be extended to weighted edges or other edge annotations, depending upon the representation required by the contextual boundary identification algorithm.
- **C**: $c_{ij} = s$ if person p_i is a member of cluster c_j, w if person p_i is weakly connected to cluster c_j, 0 otherwise. This matrix will be the output of the contextual boundary identification algorithm.
- **A**: $a_{ij} = 1$ if person p_i wrote message m_j. This matrix is assumed to be given.
- **R**: $r_{ij} = 1$ if person p_i read message m_j. This matrix is also assumed to be given.

Let $\mathbf{K} = \mathbf{R}.\mathbf{A}'.\mathbf{C}$ be similar to a matrix multiplication operation such that k_{ij} = list of messages read by person p_i which were written by authors in cluster c_j. The context and completeness hypotheses can now be tested by correlating matrices **K** and **C**. For a fixed i, there will be one x with entry k_{ix} such that $c_{ix} = s$, and many y for entries k_{iy} such that $c_{iy} = w$. The context hypothesis will imply that $u(k_{ix}) > u(\bigcup k_{iy})$ for a statistically significant fraction of people p_i, where u is the context function that calculates the average context provided by a set of messages to a recipient. The completeness hypothesis will imply that $t(k_{ix}) < t(\bigcup k_{iy} \bigcup k_{ix})$ for a statistically significant fraction of people p_i, where t is the completeness function that calculates the average completeness provided by a set of messages to a recipient.

Define the following matrix to model time evolution:

- **M**: $m_{ij} = 1$ if message m_j was written in response to message m_i, $= 0$ otherwise. **M** essentially represents a directed acyclic graph (DAG) that models time. Messages are represented as nodes, and directed edges represent a message written in response to an earlier message.

A DAG is fairly generic representation for participatory messages. Both blogs and discussions can be modeled as a DAG by considering each comment or reply as a separate message. In a blog

DAG, a comment will link back to its parent blog entry, and the blog entries will link back to other blogs or resources they may have considered. Similarly, in a discussion DAG, a reply will link back to its parent reply.

For a person p_i, a subgraph of the message DAG can now be obtained consisting of nodes of only those message that were read by p_i. Each message node can now be labeled as $\{$*strong, weak, undefined*$\}$ according to whether the messages in k_{ij} were written by a person to whom p_i had a strong or weak or undefined relationship. For each edge m_{ij} that exists in the subgraph, the following cases may arise: $s \rightarrow s$ representing further contextualization for p_i of the event under consideration, $w \rightarrow s$ representing further completeness for p_i about the event, $s \rightarrow w$ representing flow of information to an adjacent cluster, and $w \rightarrow w$ which need not be considered as relevant for p_i. This essentially models the process of information acquisition followed by person p_i. Note that in the description given earlier in Section 2.3, I only discussed the evolution of information within each cluster; but as shown here, that can be broken down into individual information acquisition processes followed by the people.

This model also allows for a more fine-grained statement of the context and completeness hypotheses. The context hypothesis should only examine whether messages from $s \rightarrow s$ helped improve the understanding of the event for a recipient, conditional on messages that the recipient has read in the past. Similarly, the completeness hypothesis should only examine whether messages from $w \rightarrow s$ helped improve the completeness of information about the event for the recipient, conditional on the messages the recipient has seen in the past. However, since the sequence in which messages were read by the participants may not be known, I have used the simpler notion of considering all the messages together, as described earlier in Chapter 2.

Appendix E

Social capital

Social capital is defined as a network characteristic emerging from the *resources* or *social assets* available with people in a social network, that are made available to other people or communities in the social network [22]. For example, the amount of trust and cohesiveness in a community that can help reach consensus on various economic and political issues, can be considered as a measure of social capital relevant in a democracy [23]. Social capital may be derived from the *structure* of networks, such as the cohesiveness between community members and their linkages with different communities. Alternatively, social capital may be derived from the *resources* available from members in the social network, such as their level of education, or their economic status, which improve the overall ability of the community for access to these resources. Researchers have also analyzed the role of social capital in poverty eradication, where the role of intermediating agencies such as civil society organizations was seen to be crucial for economic development, communal peace, and democracy [24, 25]. In fact, sociologists have recently started looking at the link between online discussions and social capital, and they have noticed significant correlations between the degree of participation in virtual and real-world interactions [26].

Appendix F

Sample surveys

F.1 Preparation of dataset: Strong and weak ties

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.

Distribution of strong and weak ties

The mass-distribution of strong and weak ties for the topics on Mumbai and books are shown in Fig. F.1 and F.2 respectively.

F.2 Hypothesis 1: Role of social ties

Strong ties of a user promote context and weak ties promote completeness in the information they receive from their topic-specific social network.

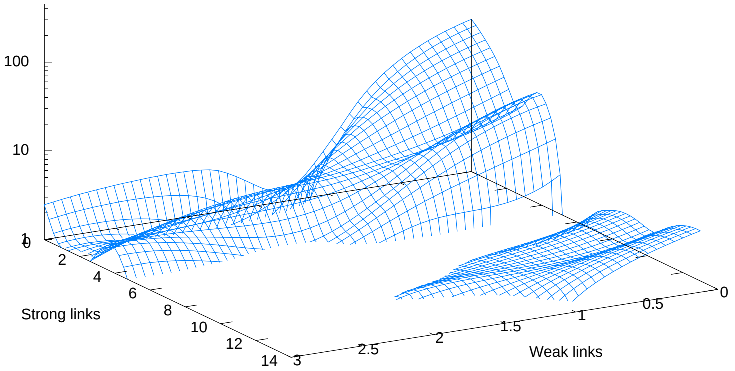

Figure F.1: Mass-distribution of strong and weak ties in the topic cluster for Mumbai

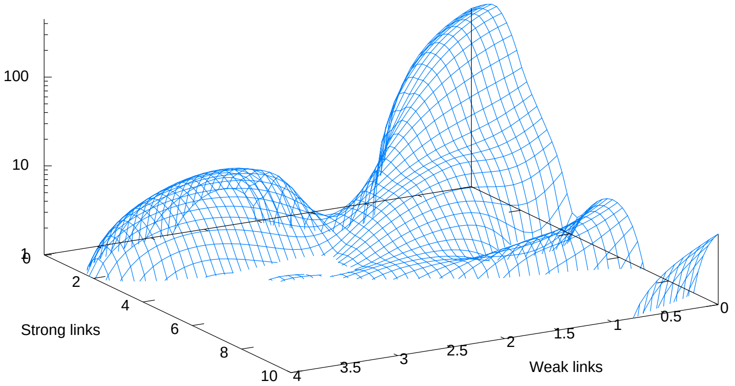

Figure F.2: Mass-distribution of strong and weak ties in the topic cluster for Books

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 F.1: Hypothesis-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 F.2: Hypothesis-1: 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 F.1, F.2, and F.3 show the results for hypothesis-1 for topics about economics, Mumbai, and books respectively.

F.3 Hypothesis 2: Informational quality of a social network

The confidence of a user to rely on members of her topic-specific social network to send contextual and complete information to her, can be measured based on the structure of her topic-specific social network.

Table F.3: Hypothesis-1: 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	$\mu = .42, n = 60$ $z = -3.79$	$\mu = .64, n = 60$ $z = -2.09$

Table F.4: Hypothesis-2: Correlation and t-tests for different topics

	Triad context	**Quad context**	**Completeness**
Economics	$\rho = .38, t = -0.37$	$\rho = .42, t = -0.14$	$\rho = .43, t = -0.09$
Mumbai	$\rho = .53, t = 0.33$	$\rho = .48, t = -0.29$	$\rho = .33, t = -0.24$
Books	$\rho = .52, t = 0.16$	$\rho = .44, t = 0.53$	$\rho = .51, t = 0.26$

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 yourself on a scale of 1-5 (1 = poor, 5 = excellent) for the following criteria:

1. How well do you feel that your friends from Orissa would be able to advice you or point you to good sources of information about development activities that you may find relevant?

2. How well do you feel your friends from Orissa are connected with other people, say friends of friends of friends... who may prove to be helpful for recommending interesting information to you related to development in Orissa?

Sample survey: Books

Seeing your interest in books, assume you had to rely on your friends to get updates about interesting books to read. Please rank yourself on a scale of 1-5 (1 = poor, 5 = excellent) for the following criteria:

1. How well do you feel that your friends who are also interested in reading, know about your particular interests in books to be able to advice you or point you to good books that you would enjoy?

2. How well do you feel your friends are connected with other people, say friends of friends of friends... who may prove to be helpful for recommending interesting books to you?

Results

Correlations between surveyed and measured values for different topics are shown in Table F.4.

F.4 Hypothesis 3: Context and completeness of messages

Context and completeness of messages can be measured based on the topic-specific social network spanned by the message.

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).

F.5 Hypothesis 4: Temporal evolution of messages

Changes with time in context and completeness of messages can be perceived by users, and correlated with measurements based on the topic-specific social network spanned by the message.

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 circumstances for people. Do you feel reply 8 on page 1 by 'Asit K' helped explain how you could use RTI in your particular circumstances?

2. Do you feel reply 1 on page 2 by 'Asit K' added some new perspectives to the use of RTI in different areas?

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 reply 5 on page 1 by 'Juhi' helped explain factors for career options that are useful for you?

2. Do you feel reply 3 on page 1 by 'a poor' added some new perspectives to career choices that can be made by different people?

Please give your ranking on a scale of 1-5 (1 = poor, 5 = excellent).

F.6 Community rankings

Sample survey: Mumbai

Seeing your interest in communities about Mumbai, please go through a few discussions in the following 5 communities and rank the communities in order of how useful you feel the discussions are for you.

1. Its Mumbai and not Bombay: *http://www.orkut...*
2. The Mumbai that I dream about: *http://www.orkut.com...*
3. Mumbai: *http://www.orkut.com/Community.aspx?cmm=...*
4. Yeh hai Mumbai meri Jaan: *http://www.orkut...*
5. Mumbai burns, yet again: *http://www.orkut.com...*

Sample survey: Books

Seeing your interest in books, please go through a few discussions in the following 5 communities and rank the communities in order of how useful you feel the discussions are for you.

1. Simply books: *http://www.orkut.com/Community.aspx?cmm=...*
2. Books: *http://www.orkut.com/Community.aspx?cmm=...*
3. Literature: *http://www.orkut.com/Community.aspx?cmm=...*
4. English literature: *http://www.orkut.com/Community...*
5. Only Good Books: *http://www.orkut.com/Community...*