# USER IMPORTANCE MODELLING IN SOCIAL INFORMATION SYSTEMS:

# AN INTERACTION BASED APPROACH

A Thesis

by

ANUPAM AGGARWAL

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2009

Major Subject: Computer Science

USER IMPORTANCE MODELLING IN SOCIAL INFORMATION SYSTEMS:

AN INTERACTION BASED APPROACH

A Thesis

by

ANUPAM AGGARWAL

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,     James Caverlee
Committee Members,   Frank Shipman
                                   Michael Pilant
Head of Department,    Valerie Taylor

December 2009

Major Subject: Computer Science

ABSTRACT

User Importance Modelling in Social Information Systems:

An Interaction Based Approach. (December 2009)

Anupam Aggarwal, B.Tech, National Institue of Technology, Kurukshetra

Chair of Advisory Committee: Dr. James Caverlee

The past few years have seen the rapid rise of all things "social" on the web from the growth of online social networks like Facebook, to real-time communication services like Twitter, to user-contributed content sites like Flickr and YouTube, to content aggregators like Digg. Beyond these popular Web 2.0 successes, the emergence of Social Information Systems is promising to fundamentally transform what information we encounter and digest, how businesses market and engage with their customers, how universities educate and train a new generation of researchers, how the government investigates terror networks, and even how political regimes interact with their citizenry. Users have moved from being passive consumers of information (via querying or browsing) to becoming active participants in the creation of data and knowledge artifacts, actively sorting, ranking, and annotating other users and artifacts.

This fundamental shift to social systems places new demands on providing dependable capabilities for knowing whom to trust and what information to trust, given the open and unregulated nature of these systems. The emergence of large-scale user participation in Social Information Systems suggests the need for the development of user-centric approaches to information quality. As a step in this direction this research proposes an interaction-based approach for modeling the notion of user importance. The interaction-based model is centered around the uniquely social aspects of these systems, by treating who communicates with whom (an interaction) as a

core building block in evaluating user importance. We first study the interaction characteristics of Twitter, one of the most buzzworthy recent Social Web successes, examining the usage statistics, growth patterns, and user interaction behavior of over 2 million participants on Twitter. We believe this is the first large-scale study of dynamic interactions on a real-world Social Information System. Based on the analysis of the interaction structure of Twitter, the second contribution of this thesis research is an exploration of approaches for measuring user importance. As part of this exploration, we study several different approaches that build on the inherent interaction-based framework of Social Information Systems. We explore this model through an experimental study over an interaction graph consisting of 800,000 nodes and about 1.9 million interaction edges. The user importance modeling approaches that we present can be applied to any Social Information System in which interactions between users can be monitored.

To my parents

## ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Dr. James Caverlee, for guiding me throughout this thesis. His constant motivation and constructive feedback helped me overcome all the technical challenges and realize my true potential. I thank him for all the inspiring lectures, for helping me in choosing my courses, for his recommendation for the Departmental assistantships, and for always being flexible with his schedule in order to accommodate my requests. I feel privileged to be one of the first Master's students to graduate under his guidance. I would also like to thank Dr. Frank Shipman and Dr. Michael Pilant for being a part of my thesis committee. I would like to acknowledge the company of all members of the TAMU infolab (Web and Distributed Information Management Lab) and in particular the members of the Twitter team. Thanks also go to my friends and colleagues and the department faculty and staff for making my stay at Texas A&M University a great experience. Last but not least, I am indebted to my parents for their love and constant support.

TABLE OF CONTENTS

CHAPTER                                                                    Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

A.   Introduction

The past few years have seen the rapid rise of all things "social" on the web – from the growth of online social networks like Facebook, to real-time communication services like Twitter, to user-contributed content sites like Flickr and YouTube, to content aggregators like Digg. Beyond these popular Web 2.0 successes, the emergence of **Social Information Systems** is promising to fundamentally transform what information we encounter and digest, how businesses and market engage with their customers, how universities educate and train a new generation of researchers, how the government investigates terror networks [11], and even how political regimes interact with their citizenry (e.g., the use of Twitter and Facebook in the recent Iranian election controversy [12]).

Unlike traditional database and web-based information systems, Social Information Systems are centered around user-contributed content, socially-generated metadata (e.g., comments, ratings, tags), and person-to-person social connections. Users have moved from being passive consumers of information (via querying or browsing) to becoming active participants in the creation of data and knowledge artifacts, actively sorting, ranking, and annotating other users and artifacts. Figure 1 illustrates how users can interact with one another in a Social Information System. Some users are content creators, some are passive consumers, while others actively participate in the content creation by tagging documents or annotating them. Users are thus, directly or indirectly in contact with one another, constantly evaluating content created

The journal model is *IEEE Transactions on Automatic Control.*
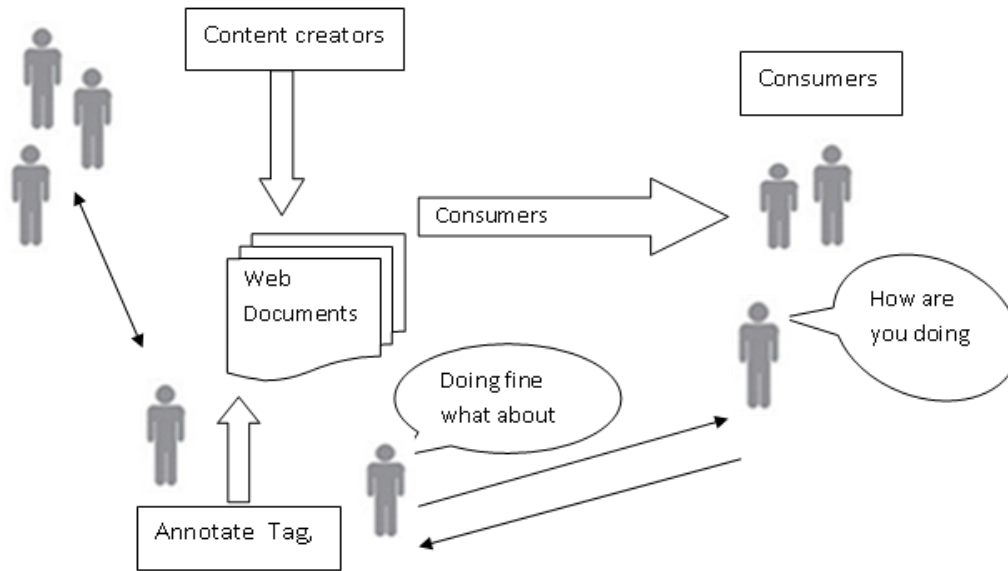
Fig. 1. Social Information System

by their peers. These new features have encouraged an explosion of user-generated content, led to the development of new modes of social information discovery, and generated a huge research interest in studying and analyzing these emerging systems.

B. Research Challenges

Along with these new opportunities, the fundamental shift to social systems places new demands on providing dependable capabilities for knowing *whom to trust* and *what information to trust*, given the open and unregulated nature of these systems. Indeed, both the database and information retrieval communities have recently recognized the immense research challenges inherent in these emerging social systems [1]. Critics of information systems incorporating social computing features argue that these systems undermine the notion of expertise: anyone can post content and share opinions even though they might not have the appropriate credentials to au-

thor the content, or might have hidden biases, leading to the content being irrelevant and misguided [30]. Thus, there is a high variance in the distribution of quality of socially-generated content.

Traditional Web approaches for assessing online information quality have typically focused on content-based and link-based approaches for assessing the quality of Web documents. Prominent examples of this style include PageRank [8], HITS [15], and linguistic analysis[6]. While these and related *document-centric* approaches have shown great success over traditional (Web 1.0) information resources (e.g., CNN, Texas A&M University, IBM), the emergence of large-scale user participation in Social Information Systems suggests the need for the development of *user-centric* approaches to information quality.

The concept of finding *importance* of users in Social Information Systems is a relatively new and encouraging research area. There have been some attempts to find high-quality user content based on the quality of comments [13], ratings of other users [16], quality of tags, and so on. However these studies have focussed more on identifying high quality content from other content (like comments and tags); they have not looked at the source of the content, i.e., the user itself. We believe that if it is possible to identify a set of important, high-quality users then content coming from them can be trusted.

A user who is important among his peers, whose opinions matter to other users on the network, and who is trusted by other users can be safely assumed to be the author of high quality content in some form, as he would have gained his following on account of his significant contributions in the past. Determining a set of good users would automatically help to ensure quality of documents, obviating the need to analyze the quality of user generated data (e.g., tags, comments, content) coming out of these users. Another reason we are interested in finding out importance is to

discover people who are popular in the network, or whose opinions matter to other users of the network. Identifying a set of important users can have several applications in advertising, and recommendation systems.

Coupled with the challenge of determining user quality is the explosion of available information about users in the system. Users leave behind a trail of online fingerprints, including the content submitted by each user, their ratings of other users and content, their annotations (e.g., tags and comments), and their connections to other users (via friendship links in the social network). While none of these individual pieces of evidence may reveal the quality of a user, it is the assertion of this thesis that these implicit signals can be harvested to support robust user importance modeling.

## C.   Overview of Thesis

As a step in the direction of modeling and assessing user quality in Social Information Systems, this thesis research proposes an *interaction-based* approach for modeling user importance. The interaction-based model is centered around the uniquely social aspects of these systems, by treating who communicates with whom (an interaction) as a core building block in evaluating user importance.

Concretely, this thesis makes two contributions:

- The first contribution of this thesis research is a large-scale study of the interaction characteristics of Twitter, one of the most buzzworthy recent Social Web successes. Twitter is a microblogging service that has attracted millions of users who communicate via short messages of 140 characters or less ("tweets"). We examine the usage statistics, growth patterns, and user interaction behavior over 2 million participants on Twitter. We believe this is the first large-scale study of dynamic interactions on a real-world Social Information System.
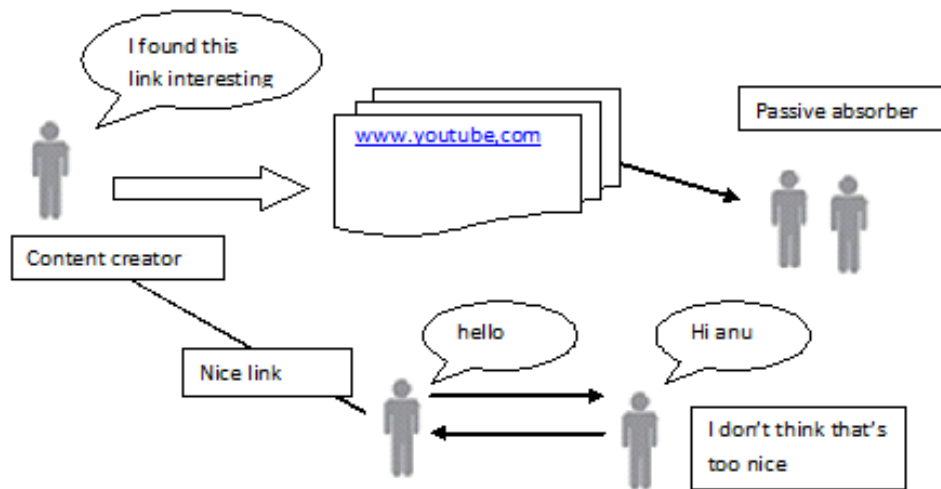
Fig. 2. Social networks: A microscopic view of Web 2.0

- Based on the analysis of the interaction structure of Twitter, the second contribution of this thesis research is an exploration of approaches for measuring user importance. As part of this exploration, we study several different approaches for measuring user importance that build on the inherent interaction-based framework of Social Information Systems. We explore this model through an experimental study over an *interaction graph* consisting of 500,000 nodes and about 1.9 million interaction edges.

While this thesis is focused on the concrete interaction patterns over the Twitter information service, the user importance modeling approaches that we present can be applied to any Social Information System in which interactions between users can be monitored. Our hope is that the insights drawn from our study of one particular system (Twitter) can be adapted and deployed in other emerging social systems.Figure 2 shows the social interactions in twitter.

The rest of this thesis is organized as follows. Chapter II discusses some of

the related research in the analysis of social networks. In chapter III we analyze the Twitter social network and describe some of the structural characteristics of the interaction patterns in Twitter. In chapter IV we present several techniques for measuring user importance, with an emphasis on approaches that leverage the interaction characteristics of the social network. Chapter V concludes the thesis by discussing some possible extensions of this research and its significance.

CHAPTER II

RELATED WORK

In this section we give an overview of related research, focused on: (i) Expert Finding and Content Quality Evaluation; (ii) Characterization of User Behavior on Twitter; and (iii) Structural Analysis of Online Social Graphs.

A.   Expert Finding and Content Quality Evaluation

Researchers have studied social interactions primarily with the aim of identifying high quality user content. In these researches about content quality user importance implicitly plays an important role. Study of interaction dynamics is undertaken by Zhang et al. [17] in their study aimed at identifying expert users in an online Java Forum, a large online help seeking community. Their aim of identifying expert users is related to our goal of identifying important people, however while their criteria of importance is mainly concerned with expert authors in social communities, our notion of importance is more generic. In our view of user importance, users don't necessarily have to be content creators, or expert contributors. They could be actors, sportsman, CEO's, politicians or anyone who is perceived as important in they eyes of society. They construct a directed post-reply graph to model interactions similar to our social interaction network and run link based algorithms similar to PageRank and HITS to calculate expertise on the underlying linked structure. Structural analysis of the java community also exhibits some similarity to ours, like the existence of bow tie structures. Campbell et al. [18] compute authority score of HITS over the user-user graph to show its correlation to content quality. Dom et al. [20] rank people on the basis of their expertise on a network of email exchanges. Agichtein et al. [16] investigate methods for exploiting community feedback to automatically identify high

quality content.

Notion of *whom to trust* is dependent upon identifying a set of *trusted* users. Our view of Importance inherently captures this notion of trust. Guha et al. [21] examine how trust is propagated in social networks by constructing a trust graph consisting of trust and distrust edges between users. Importance in our study is propagated through the web graph in a similar manner to trust.

## B.  Characterization of User Behavior on Twitter

Krishnamurthy et al. [3] have focused on studying the nature of user connections of over 100,000 profiles in the Twitter social network. As compared to our study (of over a million users), however their study is on a relatively smaller scale. Based on this study they characterize users into broadly three distinct classes: (i) *Broadcasters* or users that have a much larger number of followers then what they themselves follow (news stations like NYT, CNN etc); (ii) *Acquaintances*, who tend to exhibit reciprocity in their relationships; and (iii) *Miscreants* or *Evangelists*, users who follow a much larger number of people then they have followers. Java et al. [5] study the different types of user interactions in social networks that they have broadly classified as daily chatter, conversations, sharing information and reporting news. According to Quantcast [28] 55% of the Twitter users are female and 43% are in the age group of 18-34. Kelly et al. [27] classify tweets into five main categories. *News*, *Spam*, *Self Promotion* (tweets that promote a product/service/individual), *Pointless Babble* (which they use to classify normal tweet status messages of users, i.e., messages of type "I am going to the supermarket now"), *Conversational* (containing replies of one profile to another profile, i.e., those prefixed by an @ symbol) and *Pass-Along value* (tweets similar to forwarded emails). According to them most of the tweets are either

conversational 37.5% or in the category of Pointless Babble 40%. This fact is also observed by us in our study. Golder et al. [2] show that people message or interact with a very limited number of social connections in Facebook. They also find that message reciprocity is relatively rare even in these interactions among friends. Kumar et al. [19] characterize users as either passive members of the network; or as *inviters* who encourage offline friends and acquaintances to migrate online; and *linkers* who fully participate in the social evolution of the network.

C.   Structural Analysis of Online Social Graphs

Kumar et al. [19] study how friendship structures evolve in these social networking communities. Based on their analysis of Flickr and Yahoo 360 they segment the structure of these social networks into three regions: singletons who do not participate in the network; isolated communities which overwhelmingly display star structure; and a giant component anchored by a well-connected core region which persists even in the absence of stars. Our analysis of the interaction structure of Twitter yields very similar results. We also find the existence of a large strongly connected component formed as a result of profile interactions along-with the presence of large number of singleton profiles who do not participate in the community.

In their study of Twitter, Krishnamurthy et al. [3] have observed a correlation between tweet frequency of the user and the number of followers of the profile. This correlation between tweet frequency of a user and his follower/following count is the motivation for our idea of biasing PageRank computation by the number of follower/followers. Most of their findings about user behavior, social relationships are consistent with other earlier related studies [5], however their study of friendship based interactions is based, just like other related studies, on examining the

immediate profile neighborhood. They only look at direct relationships (or just the immediate followers and following profiles). Our approach on the other hand is much more global in sense that it looks at the global graph of such friend relationships and their interactions.

Java et al. [5] do some usage analysis which shows that the growth rate of Twitter has slowed possibly because the initial hype has died down, along with the fact that user activity declines with time for a vast majority of users suggesting that there are a sizeable amount of users who join the networks out of curiosity. This fact also underlines that a traditional friendship based approach for detecting user importance is not suitable as it does not capture user activity over time. Their study of social networking relationships is also, like earlier studies based on the traditional friendship (follower-following) graph structure. They construct a directed graph G(V,E) where V represents the users and E(edges) represent the set of friend relations where a directed edge e between two users u and v exists if user u declares v as a friend. Their analysis of the nature of this friendship graph and its properties like average degree of a node, reciprocity, existence of strongly connected components, size, etc. also indicate that the network shows a large amount of reciprocity in the friend graph suggesting that new friends join the network mainly by invitation from other friends. This leads us to believe that one aspect of measuring user importance can be simply the number of friends a profile has (which is the motivation for biasing TrustRank computation by the number of followers). However this is not entirely true as there are users who follow updates of large number of other profiles (thus becoming their friends) for dubious reasons (for example a recruiter can follow updates of large number of profiles to find out about profiles that are looking for networking opportunities, or employers can follow updates of their employees to keep tabs on their activities).

CHAPTER III

ANALYSIS OF INTERACTION STRUCTURE ON TWITTER

The first contribution of this thesis research is an analysis of the interaction structure of Twitter. In the past much of the study of social networking communities has focused around the static relationship structure among participants, e.g., the declared friendship structure of Facebook members. However a study based on friendship-based linked structure does not reveal the actual interactions among people. Such a study does not reveal, for instance, the friends people communicate more often with and who among them reciprocate this attention. Golder et al. [2] show that people message or interact with a very limited number of social connections in Facebook. They also find that message reciprocity is relatively rare even in these interactions among friends.

As part of this first research thrust, we investigate the dynamic interactions on a real-world Social Information System. In particular, we study the characteristics of one such social network Twitter, its usage statistics, growth patterns and user interaction behavior. We study several standard metrics like average indegree and outdegree of a node, the degree of reciprocity, the percentage of terminal and dangling nodes in the network, and so on. We believe this is the first large-scale study of dynamic interactions on a real-world Social Information System. The results of this study have important implications for modeling user importance, for example, by considering the frequency of interaction between individuals as an approximation of their relationship strength. In the rest of this section, we present an overview of Twitter, describe the Twitter data collected as part of this study, discuss how we can view Twitter as a social interaction network, and present an analysis of the interaction structure on Twitter.

A. Twitter: An Overview

Twitter is a free social networking and micro-blogging service created by Jack Dorsey in 2006, that enables its users to send and read short messages of 140 characters known as *tweets* [23]. These tweets are delivered as feeds to profile pages of all the *followers* of that profile. Followers are people who subscribe to receiving the person's updates. Senders can either make their feeds *private* in which case they can only be viewed by their friends (or followers) or make them *public* in which case they can be viewed by anyone on the network. Twitter has a broadcast nature of message delivery which is one of the reasons why it has become such an important source for discovering real time events. Users can send and receive these tweets not only through the website but can also use their phones, and other third party applications to tweet. According to Quantcast.com [28] Twitter had about 30 Million profiles as of July 2009.

Figure 2 gives an overview of user interactions in Twitter. In Twitter people can either interact directly or indirectly. They can post links to video's, photos or other content, inviting comments from other users. Twitter helps people to keep tabs on the activities of set of people they think are important.

Table I. User composition and popularity of Twitter [26]

| Country | Percentage of Users | Traffic Rank in the Country |
|---|---|---|
| USA | 42.5 | 13 |
| Germany | 8.1 | 14 |
| India | 7.3 | 13 |
| UK | 6.2 | 13 |
| Japan | 2.9 | 63 |

Twitter has been described as the SMS of the internet by [25]. It is one of the fastest growing (growth rate of 1382% per month) and amongst the top 50 most popular websites of the world[26]. Most of the users on Twitter come from the United States. The detailed composition is given by Table I. Essential to the real time characteristic of Twitter is a real time search engine which can instantly index any public update sent from anywhere in the world. This enables Twitter to keep track of happenings around the world and function as a real-time discovery engine of news from all over the world.

Figure 3 shows an example of a profile on Twitter for user Ashton Kutcher. Ashton Kutcher is amongst the most popular profile on Twitter with 3,345,264 followers following his feeds. He in turn follows 207 other profiles. Out of the several tweets on his page the tweets prefixed by an @ sign are the replies from aplusk's profiles to other users. These replies are usually addressed to a single profile and can be viewed as an indication of direct communication intent between two users. The other tweets (those not prefixed by an @ sign) can be general statements or comments targeted at the entire follower population.

B. Dataset Description

We model our approach of detecting user importance using data crawled from the Twitter social network. Our data collection methodology involves crawling the public timeline of Twitter [14]. The Public Timeline of Twitter is a profile page consisting of tweets sampled randomly from all the public profiles on Twitter. This profile provides all the data for our study [14]. Crawling the public timeline can give a fairly accurate snapshot of events going in the world at that particular time. We crawl the public timeline and gather about 600 tweets per minute. The tweets are stored in xml format

Fig. 3. Twitter profile of Ashton Kutcher (aplusk)

with roughly 15 days of tweets being used for analysis at a time.

```
−<status>
    <created_at>Mon Jan 19 18:53:42 +0000 2009</created_at>
    <id>1131110627</id>
  −<text>
      @susanmccool LOL - thanks when I can afford you I'll hire you to manage my email. :)
    </text>
    <source>web</source>
    <truncated>false</truncated>
    <in_reply_to_status_id>1131100510</in_reply_to_status_id>
    <in_reply_to_user_id>15235214</in_reply_to_user_id>
    <favorited>false</favorited>
  −<user>
      <id>14402079</id>
      <name>Lamar Romero</name>
      <screen_name>FundingUrBiz</screen_name>
    −<description>
        Devoted husband, entrepreneur, servant leader. interests: Getting businesses funded! Dis
      </description>
      <location>Austin</location>
    −<profile_image_url>
        http://s3.amazonaws.com/twitter_production/profile_images/52846090/Lamar_in_white
      </profile_image_url>
      <url>http://inaustin.ning.com/</url>
      <protected>false</protected>
      <followers_count>44</followers_count>
    </user>
</status>
```

Fig. 4. xml format of the stored information crawled from public timeline

Figure 4 shows the xml format in which we store the information crawled from the public timeline. The tweet contents are stored between the *text* nodes. The user information corresponding to that tweet is contained in the *user* nodes. This xml snippet consists of information about a user named Lamar Romero having the screen name of FundingUrBiz. The text node contains his reply to another profile having the name of susanmcool.

This 15 days of public timeline data contains information about roughly 2 million distinct users. Since the public timeline is cached for 60 seconds we hit the public timeline page once per minute, each time collecting about 100 tweets.

Table II shows that the number of distinct users found out from a crawl of the

Table II. Basic usage statistics of Twitter dataset

| Statistic | 1-17 Dec 08 | 17-31 Dec 08 | 31 Dec-19 Jan 09 |
|---|---|---|---|
| No. of Users | $1,171,631$ | 1,075,142 | 1,433,158 |
| No of Users receiving replies | $393,502$ | 346,428 | 468,746 |
| No of users posting general comments | $698,795$ | 649,903 | 864,548 |
| No. of RepliesTo | $3,662,112$ | 3,357,671 | 4,625,594 |
| No. of Total Tweets | $12,509,510$ | 11,734,153 | 15,240,176 |

public timeline remains fairly constant (around 1.5 million for 15 days). Approximately 25% of the tweets are replies to, the other being general comments. This result is consistent with the trend observed by [27].

Figure 5 shows the distribution of the frequency of replies to profiles. This graph gives us an idea of the nature of tweets of users. The majority of users have fewer than 100 replies addressed to them. This shows that users primarily do not use replies addressed to their individual friends to communicate. This behavior can be attributed to the broadcast nature of Twitter by which user status feeds are visible to all of their friends. Since the number of replies are scarce we can safely assume that the users who are the addresses of the replies are important to the person replying directly to them. This forms the basis of our technique of modeling replies as a vote of importance or as an intent of showing special interest to the receiving user.

C.   Viewing Twitter as a Social Interaction Network

In this section, we show how to view Twitter as a social interaction network. We begin our study by first considering the *declared social network* corresponding to the declared relationships in the network (e.g., followers, following). This static network gives only a partial view of the dynamics of the *interactions* that make a social network vibrant; hence, we then study the interaction dynamics and usage statistics of Twitter. This is essential in order to understand the nature of communication
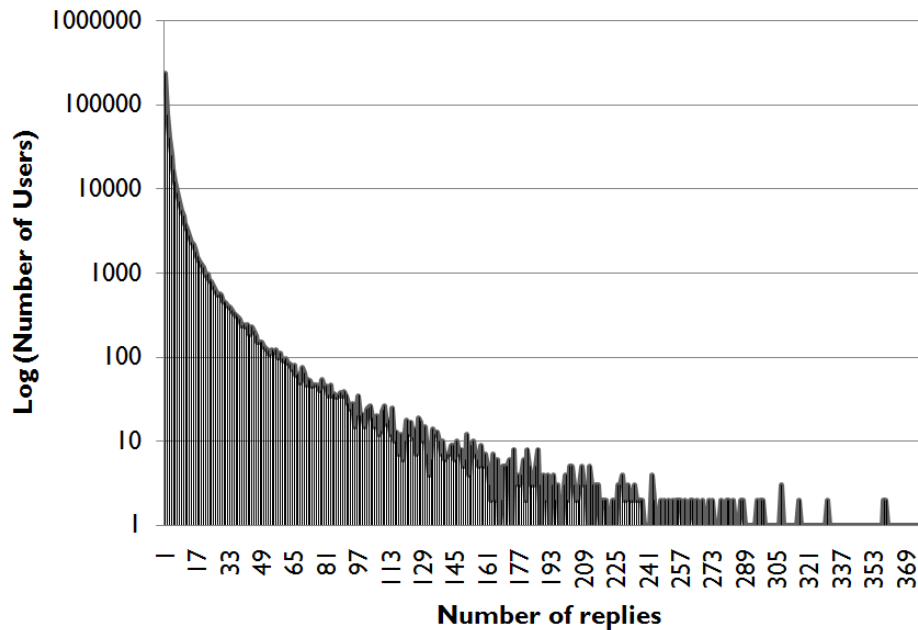
Fig. 5. Distribution of reply frequencies with profiles

between Twitter users. This would help us answer questions like, how do users in Twitter interact with one another, what is the primary interface they use, what is the frequency of these interactions, what is the nature of most of the tweets, whether they are general comments or addressed to friends directly etc.

## 1. The Declared Social Network

A *declared social network* is constructed over the following relations in Twitter. Following profiles are all those profiles whose updates a profile subscribes to receiving. They indirectly model expression of interest in the profile being followed as typically users follow all those profiles which are of interest to them. These profiles might be of their close friends or some social celebrities. Thus these relations indirectly model notion of profile importance as perceived by the followers.

We constructed the *declared social network* over all the profiles in our dataset.
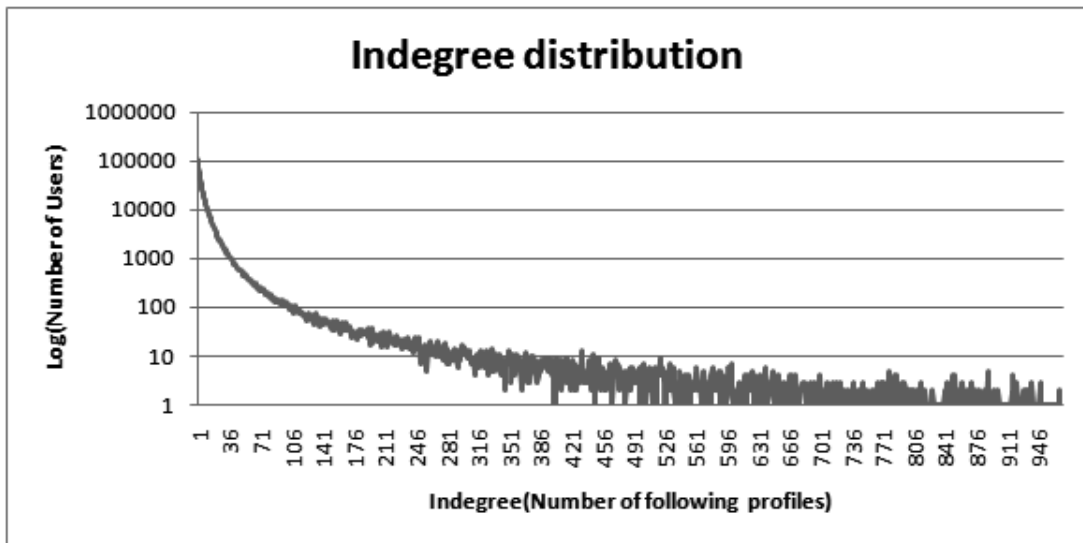
Fig. 6. Indegree distribution of following graph

We also discarded all those profiles that were being followed but were not in our dataset. Formally a *declared social network* is a graph constructed over following relations is a directed graph G(V,E) where the vertices V represent the profiles and and E represent the directed edges. If a profile A follows profile B there would be an edge from A to B. E thus represents the entire set of these edges.

The reason why we constructed a graph over the following relations and not the follower relations is because of the fact that following profiles are less in number whereas the followers might me of order of several thousands. This makes graph construction easy as it takes less memory and computation is more efficient. However since both follower and following relations are reciprocal it really does not matter whether graph is constructed over the followers or the following.

Figure 6 shows the indegree distribution of the friendship graph and figure 7 shows the outdegree distribution. The indegree distribution shows an exponential decrease in the number of users with increase in indegree. Indegree of a profile can be indirectly correlated to the number of followers of a profile. A profile having
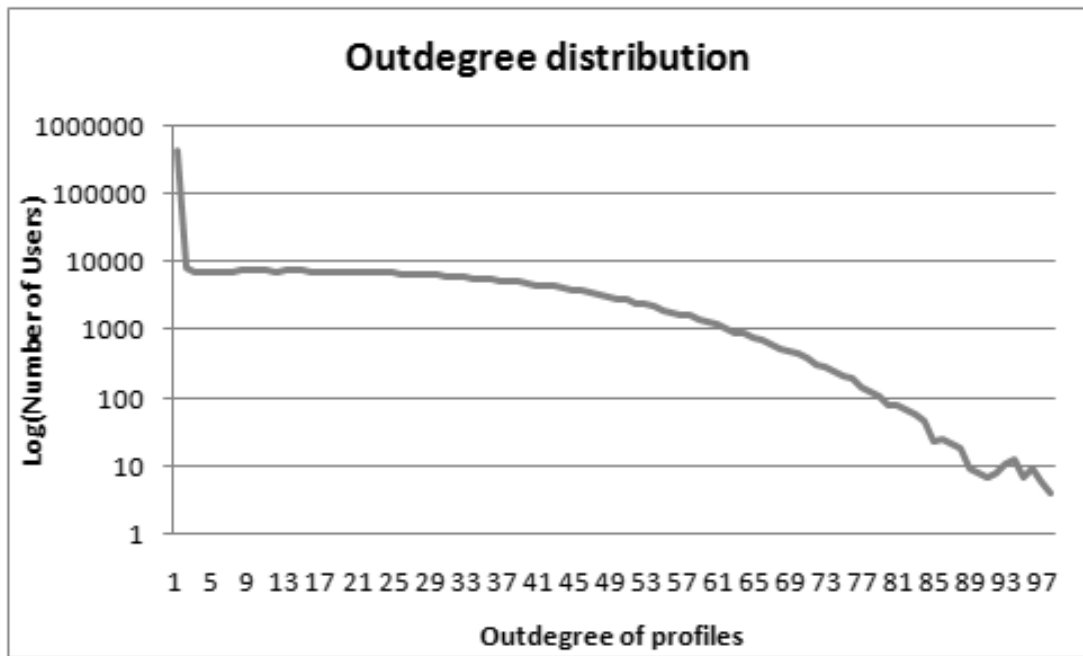
Fig. 7. Outdegree distribution of following graph

large indegree would have a large number of followers and vice versa. The outdegree distribution however gives an indication of the number of profiles followed by the profile in consideration. From the graphs we see that the number of profiles becomes very less as outdegree increases beyond 100. This goes to show that more than 90% of users on twitter follow less than 100 profiles. In contrast however there are still a substantial number of users having indegree more than 100.

Table III gives some of the statistics of the friendship graph. In the graph the maximum indegree profile is of Barack Obama. Maximum indegree node signifies a profile having the maximum number of followers in our dataset. Barack Obama is the current president of US and one of the most important persons in the world. This goes to show that the indegree or the number of followers of a profile can be a fairly useful metric for measuring profile importance.

Table III. Friendship graph statistics (constructed over same dataset as interaction network)

| Statistic | Value |
|---|---|
| Number of Nodes | 802144 |
| Average Indegree and Outdegree | 10.7 |
| Max Outdegree | 97 |
| Max Outdegree Profile | speakeroftruth |
| Max Indegree | 34199 |
| Max Indegree Profile | BarackObama |
| Percentage of dangling nodes | 57 |

## 2. The Interaction-Based Social Network

Since Twitter users can communicate with other users through the use of the "reply" mechanism, regardless of declared friend relationships, we can view these direct replies as interactions. In the aggregate, we view the collection of users and their replies as the basis of a social interaction network. Formally, a *social interaction network* is a directed graph G(V,E) formed over the replies where V represent the nodes in the graph which represent user profiles and E represents the directed edges indicating the reply. A directed edge from user A to B indicates that at some point user A tweeted user B (e.g., A tweeted a message to B in the form "@B...message..."). We construct two variants of this Reply-to graph the *weighted* variant and the *unweighted* variant. In the weighted variant of the social interaction network, the edge weights on the edges indicate the frequency of the interaction between users representing the edge nodes. If profile A tweets profile B 5 times we put an edge weight of 5 on the edge A-B.
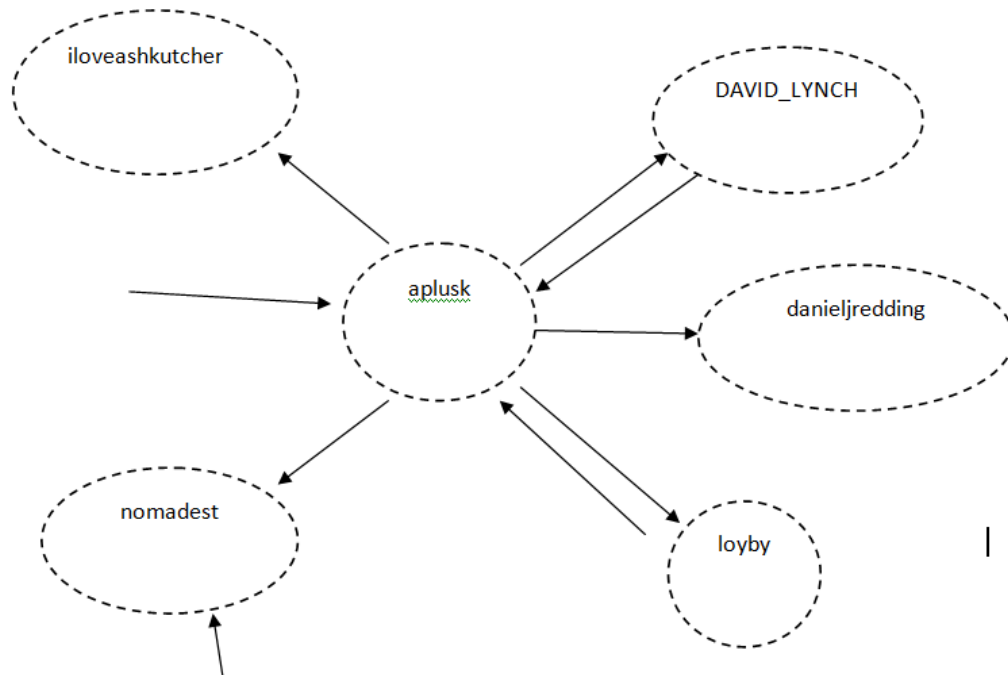
Fig. 8. Social Interaction Network in neighborhood of profile aplusk

As an illustration, Figure 3 shows a snapshot of Twitter user aplusk (Ashton Kutcher) and Figure 8 shows the corresponding social interaction network in the immediate neighborhood of aplusk. Unlike the declared social network, which is a static graph not affected by the extent of communication between friends, the interaction network captures the frequency of interactions between the users on twitter. It thus gives an indication of profile activity of a profile. Typically a profile which uses twitter very frequently would have a dense network of edges around it. Disinterested/infrequent users therefore can be easily found out using this graph.

Table IV gives an overview of the properties of the social interaction network constructed over a period of 15 days. The graph consists of lots of dangling nodes. These are profiles that do not participate in conversation with anyone but might have incoming tweets directed at them. These properties of the social interaction

Table IV. Reply-to graph statistics (constructed over 15 days period)

| Statistic | Value |
| --- | --- |
| Number of Nodes | 802144 |
| Average Indegree and Outdegree | 2.1 |
| Max Outdegree | 934 |
| Max Outdegree Profile | Mstweet |
| Max Indegree | 7661 |
| Max Indegree Profile | stephenfry |
| Percentage of dangling nodes | 32 |

network are fairly consistent across time (for a graph constructed over roughly the same duration of time in the weeks immediately preceding and following the period of study).

We observe some basic differences in the interaction graph from the declared friend graph on examining Table IV. One of the most important difference is the average indegree and outdegree. This is due to the fact that in a declared friend graph a profile typically follows many other profiles. An interaction graph only captures profiles these profiles actually interact with. From the value of indegree and outdegree it is evident that profiles only interact with a subset of their friends. Also the maximum outdegree of profile in the social interaction network is a lot more than the friendship graph. This indicates that there is large variance of user characteristics in the social interaction network with some users tweeting others a lot. A declared friendship graph on the other hand does not have these large variances in the degree distribution. This variation is also evident after comparing the outdegree distribution of the social interaction network and the declared friend graphs (Figures 10 and 7).
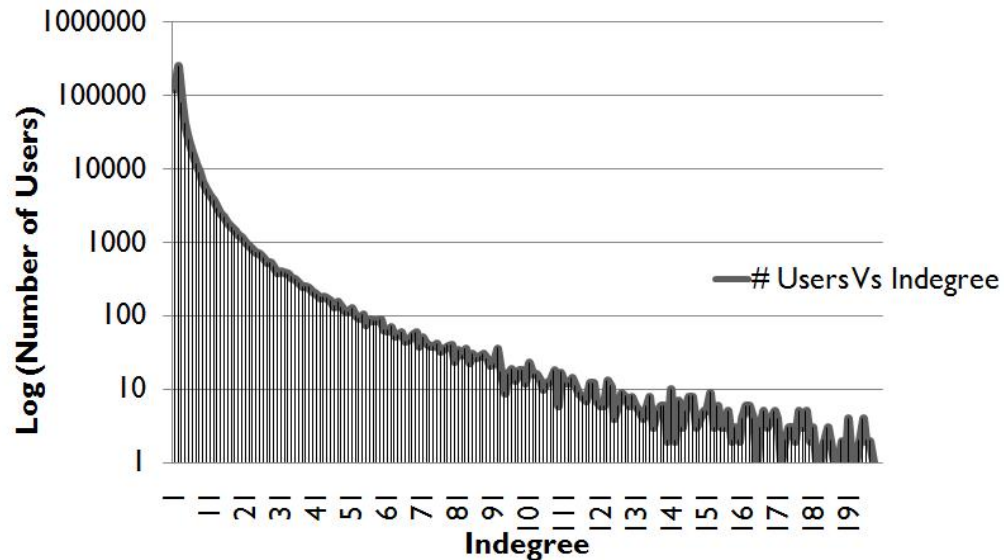
Fig. 9. Indegree distribution of nodes in reply-to graph

One interesting statistic is the maximum outdegree of a node and the corresponding user profile on Twitter. A node having the maximum outdegree on the social interaction network will represent a profile that has tweeted a lot of other profiles. If the number of outlinks from this profile is unusually large than there are high chances of the profile being a spam profile. This holds true for the maximum outdegree node profile in our social interaction network also. In our graph we find that the profile MStweet corresponding to the node with the highest outdegree is indeed a spam profile that has been suspended by Twitter. This suggests that a node having a large number of inlinks (The maxindegree profile is that of stephenfry who is a famous writer, blogger from UK) can be safely assumed to be an important profile. While this assumption holds true generally it is not the sole criteria for a profile to be important.

Figures 9 and 10 show the distribution of indegree and outdegree of nodes in the Twitter network.
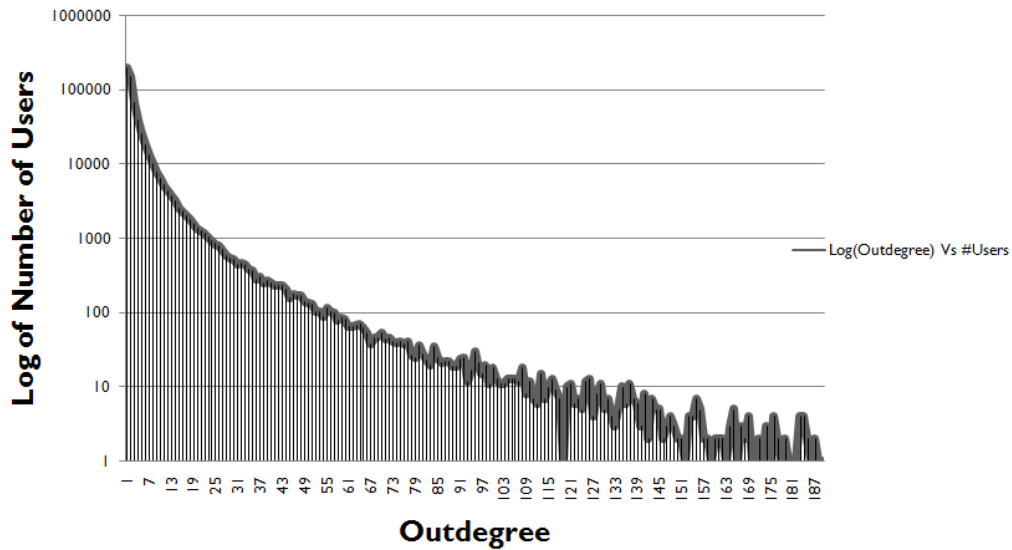
Fig. 10. Outdegree distribution of nodes

The distribution of indegree and outdegree for the social interaction network are very similar. This is expected as the indegree and outdegrees are reciprocal in nature. The number of users decrease exponentially with increasing number of replies. Most of the users in this graph have a low indegree/outdegree which indicates that within the 15 day period majority of users are not very active. They usually converse with less than 25 different friends. The graph therefore helps us to find those few important friends whom majority of users talk to. In comparison to the declared friend graph the decrease in the outdegree/indegree is much higher in the social interaction network.

We also observe strongly connected components of varying sizes in the social interaction network. The strongly connected components are a set of maximal strongly connected subgraphs. There is a path from each vertex to every other vertex in such components. In our context they can be viewed as a set of communication chains that connects different profiles. The strongly connected components were calculated using Tarjan's algorithm. A regular feature observed in the social interaction networks

Table V. Strongly connected components (for Feb 1-19th Dataset)

| Number of SCC's | Size |
| --- | --- |
| 9504 | 2 |
| 1594 | 3 |
| 492 | 4 |
| 199 | 5 |
| 92 | 6 |
| 39 | 7 |
| 23 | 8 |
| 16 | 9 |
| 7 | 10 |
| 5 | 11 |
| 5 | 12 |
| 1 | 13 |
| 2 | 14 |
| 1 | 17 |
| 1 | 16 |
| 1 | 220638 |

constructed over different period was the existence of one big strongly connected component.

In Table V we observe one component of size 220,638. The regular existence of this big strongly connected component can be attributed to the crawl methodology of the Twitter Public timeline. Public time line crawl typically goes breadth wise from some select random profiles. Because of this breadth wise crawl typically follower and following profiles of the seed profiles also get crawled. Since most of the replies are addressed to the followers and the following profiles the graph formed tends to have a connected nature.

Table V shows the number of strongly connected components of each size. The composition of some of these components are shown in Table VI. Table VI shows the existence of *bow-tie* structures in the social interaction network. Presence of bow-tie structures is a regular feature in the webgraph. Figure 11 shows what a bow-tie

Table VI. Composition of strongly connected components

| Mutual links | Inlinks(Links into component) | Outlinks | Size |
|---|---|---|---|
| 1339588 | 106497 | 291538 | 220638 |
| 72 | 5 | 12 | 17 |
| 61 | 3 | 10 | 16 |
| 37 | 1 | 3 | 14 |
| 28 | 13 | 12 | 13 |
| 24 | 1 | 20 | 12 |
| 18 | 9 | 25 | 11 |
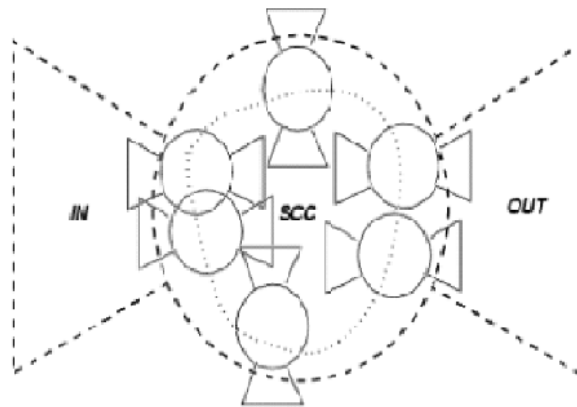


Fig. 11. Bow-tie structure

structure is. The social interaction network has many characteristics similar to a webgraph constructed over hyperlinks, this suggests that we can use link structure based algorithms such as PageRank that determine importance of web pages to model user importance in these social interaction networks.

In this section we looked at the interaction characteristics of users on Twitter. We examined some usage characteristics of Twitter and gave an overview of our data collection methodology. As part of this study, we described the construction of the *declared social network* and the *social interaction network* and studied the properties of both graphs. This also gives us important insights about user behavior and their

interaction characteristics. Based on this analysis, we propose in the following chapter several approaches for measuring user importance in a Social Information System.

CHAPTER IV

EXPLORING USER IMPORTANCE MEASURES

Based on the analysis of the interaction structure of Twitter, the second contribution of this thesis research is an exploration of approaches for measuring user importance. As part of this exploration, we study several different approaches for measuring user importance that build on the inherent interaction-based framework of Social Information Systems. We explore this model through an experimental study over an *interaction graph* consisting of 800,000 nodes and about 1.9 million interaction edges.

Determining a set of important users has applications in advertising, recommendation systems, and in improving the quality of content in social networks on the whole. It can help us to determine whom to trust, and consequently what information to trust. In this section we propose a set of methods to examine this notion of user importance. Our notion of user importance includes all people who gain importance in social networks by virtue of their contribution to the community (e.g., good sources of comments, tags), and also on account of their celebrity status (like politicians, athletes, actors, CEOs, etc.).

A. Modeling User Importance: Initial Approaches and Results

In this section we examine some basic approaches to model user importance and examine some of their limitations along with analyzing the results obtained by following these approaches. Note that there is no gold standard for whether the users identified by the proposed approaches are indeed "important". Hence, this chapter serves primarily as an exploration of several approaches; future research will be needed to evaluate the user quality estimates through a formal user study or application scenario. Also, our study of user importance is dependent on how users communicate in

the social network and how direct interactions are represented in the social network. In our study for example we parsed @ prefixed in front of the profile names in the public timeline to extract the direct messages. In Facebook this convention is not used for direct messages, instead users comment on each other's statuses directly under the message text. Though our approach for modeling importance based on direct interactions is general to most social information systems, how those direct interactions are captured depends on the characteristics of the social networks in consideration. It is also affected by how users use the social network. For example in Twitter,a small percentage of people might not prefix @ in front of the receiving users profile name to communicate with them. In Facebook they might send direct messages or emails instead of commenting on the status.

## 1.    Number of Followers

One simple criteria for measuring the importance of a profile on Twitter can be the number of followers of the profile. Twitter has a lot of celebrities, along with lots of fans who follow their updates. A *celebrity* profile can be any profile that is perceived as important in the society by other peer profiles. They include profiles associated with entertainment industry, sports, politics or any other walk of life.

The broadcast nature of Twitter allows people to keep tabs on the activities of these celebrity profiles. In our study we have observed that most of these celebrities have a large number of followers. On compiling a list of all profiles by the decreasing number of their follower counts we observe that several socially recognized celebrities figure in the list. Table VII shows some top profiles by the follower count.

The results show that even a basic metric like follower count can be useful in determining importance. However the limitation of this method is that it is Twitter-specific as there is no concept of followers, in other social networks. Moreover, this

Table VII. Top profiles by follower count for Jan 31-Feb 19th dataset

| Profiles | Follower Count | Rank | Description |
|---|---|---|---|
| britneyspears | 274406 | 1 | Internationally renowned singer |
| stephenfry | 272708 | 2 | Famous writer, British actor & blogger |
| cnnbrk | 246423 | 3 | Profile of world famous news network CNN |
| nprpolitics | 210459 | 4 | Profile of NPR news |
| mashable | 209238 | 5 | Popular Internet news blog Mashable |
| kevinrose | 197471 | 6 | Founder of social media site Digg |
| lancearmstrong | 195822 | 7 | Seven times winner of Tour de France |
| THE_REAL_SHAQ | 193886 | 8 | Famous NBA player Shaquile o Neal |
| Twitter | 191080 | 9 | Profile operated by Twitter itself |
| ev | 180960 | 10 | Profile of CEO of Twitter, Evan Williams |
| aplusk | 179463 | 11 | Famous Hollywood actor Ashton Kutcher |
| MCHammer | 179090 | 12 | Rapper, entertainer and dancer Hammer |
| TechCrunch | 175134 | 13 | Weblog profiling new internet products |

method helps us to discover only socially recognized celebrities. This method for instance does not help to discover quality content contributors, or people who might not be popular in society as a whole but are very popular among their friends. After examination of Table VII we see that all the profiles are very well recognized. There is a good mix of influential people from all walks of life: actors, news and entertainment, bloggers, CEO's etc.

Another limitation of this method is that it does not capture user activity. Lots of people join social networks like Twitter just because of the initial hype; after a while they do not participate as much and consequently might lose their importance. A method based on follower counts would not be able to take into account this message dynamics and profile activity. However one advantage of this method is that it is reasonably robust to spam. This is because spam profiles on Twitter typically do not have so many followers, hence they would show up low on the list of top follower users.

## 2.   Number of Incoming Tweets

Another criteria for measuring the importance of a profile can be the number of incoming messages to a profile. The number of incoming messages indicates interest in a profile. The basic motivation for this approach is that people usually interact with their close friends through direct messages, rather than general status updates meant for the consumption of all the followers. The Twitter equivalent of these direct messages are messages prefixed by the @ sign. Counting the number of incoming tweets into a profile can thus serve as an important indicator of importance.

We rank users in decreasing rank by the total number of tweets directed at them (which might include multiple tweets from the same user) and also by the number of unique users tweeting them. Table VIII shows the top profiles by the total number of tweets for the duration of Jan 31-Feb 19th(counting distinct tweets from same profile as different), whereas Table IX shows the top profiles by the number of unique users replying to that profiles (only one tweet is counted per user).

After analyzing these results we observe some interesting trends. After comparing Table IX and VIII, the results of Table IX seem more intuitive. In the case of top users by total number of tweets a single profile can be responsible for a large fraction of the tweets received by its neighbors. This is typically the case with spam profiles or with profiles which have a habit of tweeting others profiles excessively. In Table VIII we do observe some spam/suspicious profiles. In fact the profile having a rank of 1 turned out to be a spam profile. While ideally a spam profile should have lots of outgoing messages it can also show up high on the list of top tweet receivers as its spam tweets might prompt other users to reply to that particular profile resulting in a high incoming tweet count. Table VIII contains profiles of some regular users for which we did not find any indicators of profile importance after examining the profiles.

Table VIII. Top profiles by total incoming Tweet frequency for Jan 31-Feb 19th dataset

| Profiles | Rank | Description |
|---|---|---|
| masterconsole | 1 | Account suspended due to suspicious activity(Bot) |
| badchewy | 2 | Self described "search" guy/passionate about search |
| reasio | 3 | Account no longer exists |
| Umma | 4 | Regular user profile(No special importance indicators) |
| Lady12s | 5 | Regular user profile(No special importance indicators) |
| iamtheplague | 6 | Self described Twitter addict |
| HerrTwiggs | 7 | Locally popular musician based in Germany |
| eljuncoenpie | 8 | Regular user profile |
| al67 | 9 | Avid guitarist and repairs guitar(Regular user) |
| rdarmanin | 10 | Wannabe journalist and blog writer |
| KmD25 | 11 | Suspicios user/spam profile |
| erinwarde | 12 | Blogwriter |
| kaboogie | 13 | Owns small time recycled leather business |
| Takammy | 14 | Regular user |
| mountflorida | 15 | Writer, Blogs about career in Music |

It could just be that these users received lots of tweets from a relatively smaller subset of their friends during this interval which resulted in their high rank on the list. This is evident also from the fact that some of the users in our table describe themselves as Twitter addicts, they might be using Twitter a lot to communicate with their select friends which might lead to a high overall incoming Twitter count.

We also observe that some profiles which figure in this list have changed significantly since the time their activity was last captured by our dataset. One such example is of profile kmd25. On examining this profile we do not see any indications of its high rank (the profile has 1 follower and does not follow any other profile). This leads us to believe that this profile is most likely a Bot or a spammer as if it was a genuine profile it would have had more followers replying to this profile.

However importance as measured by count of unique users replying does seem to be a good measure of finding importance. All the profiles are genuine and of relatively

Table IX. Top profiles by number of unique users Tweeting for Jan 31-Feb 19th dataset

| Profiles | Rank | Description |
|----------|------|-------------|
| stephenfry | 1 | Famous writer, British actor |
| Schofe | 2 | Profile of Philip Schofield, a British television presenter |
| ijustine | 3 | Avid Twitter and blogger |
| bobbyllew | 4 | Robert Llewellyn's profile-English actor, presenter, and writer. |
| lilyroseallen | 5 | Lily Allen's profile-English recording artist, talk show host and actress |
| ricksanchezcnn | 6 | Profile of Rick Sanchez, a CNN news anchor |
| warrenellis | 7 | A writer based in England |
| xxandip | 8 | Profile of Andy Peters, a TV personality |
| feliciaday | 9 | Felicia day's profile-she is an actress, gamer and Misanthrope |
| Fearnecotton | 10 | Profile of Fearne Cotton a popular British TV presenter |
| chrispirillo | 11 | Founder and maintainer of network of blogs, web forums |
| greggrunberg | 12 | An American TV actor |
| LeoLaporte | 13 | US based technology journalist |

important personalities, as is indicated from Table IX. One more interesting aspect of this result is that we find lots of writers and bloggers on the list. This method helps us to capture influential content creators of social media. Hence it gives us a different flavor of importance than the follower count based method. These profiles although not very important in society on the whole, are profiles of individuals who contribute a lot to social media and profiles other people communicate with frequently.

This result is according to our expectations as typically a writer or a blogger would have a tendency to tweet more than average as compared to a regular Twitter user. This leads to these profiles attracting more replies as well resulting in high tweet count and rank. We have also observed a trend in Twitter these days of ordinary people reaching out to these writers, journalists directly by means of tweets (addressing them using @) in order to post their views about some article that the blogger or writer wrote.

From the above discussion we conclude that rank by number of unique users tweeting the profile is a better method than rank by total number of incoming tweets which does not capture notion of importance as effectively. In the following approaches we combine both the factors of measuring importance by profile activity (by assigning weights according to frequency of interactions) and importance perceived as by peer users (based on follower count).

B. Modeling User Importance: Using the Social Interaction Network

In this section we describe some of the approaches to measure user importance using the social interaction network. These random walk approaches have their origins in citation analysis developed in the 1950s by Eugene Garfield at the University of Pennsylvania, a concept borrowed by Larry Page et al. in their influential PageRank paper [8].

### 1. Random Walk Importance Model

The first approach assumes that a profile importance can be gauged from not only the aggregate number of inlinks to the profile in the social interaction network but by also taking into account which profiles those links come from. Thus, we look at a profile's importance from the perspective of the importance of its peers, similar to how a web page importance is determined by the importance of the webpages having outlinks to it. Thus for our social interaction network constructed in the neighborhood of Ashton Kutcher (Figure 3) if Ashton Kutcher who is an important celebrity tweets users David_lynch and loyby some of the importance of Ashton Kutcher would be transferred to david lynch and loyby.

Formally, the importance of a user profile can be stated as:

$$Imp_u = c \sum_{v \in B(u)} Imp_v / Nv \qquad (4.1)$$

where u represents a user profile. B(u) is the set of profiles that tweet u. *Imp(u)* and *Imp(v)* are importance scores of profiles u and v, respectively. *Nv* denotes the number of outgoing tweets of profile v and c is a factor used for normalization.

or in matrix form

$$Imp = \alpha.T.Imp + (1 - \alpha).1/N.I_N \qquad (4.2)$$

where *Imp* is the user importance vector and T is the transition matrix.

In a random surfer model over the social interaction network the movement of the random walk is determined by the transition matrix T. The surfer can either follow one of the outlinks from a profile or if a profile does not have any outlinks randomly choose any profile to move to. To take into account this random teleportation to a random profile we choose a parameter $\alpha$ having a value of 0.85. This basically means that the surfer is guided by the transition matrix 85% of the time and 15% of the time he randomly picks up a profile to visit. The matrix T has a property that all the entries are non negative with all the entries in each column summing upto 1, hence it is a stochastic matrix.

The matrix Imp is a stochastic vector as it is a combination of stochastic matrices, implying it has a stationary state. The rate of convergence to steady state can be controlled through modifying the parameter $\alpha$. After sufficient iterations the Importance vector converges to steady state after which the importance scores do not change.

Intuitively, the random walk method is better than the static follower count method because it is immune to user inactivity. It is also better than the tweet

frequency method because it does not look at the immediate followers/following behavior, but instead looks at the global behavior of all the users on the network. The other advantage of the random walk model is that it also considers importance of friends in order to calculate profile importance. In this way profiles who are friends with important people would be more important than profiles who are friends with less important people.

Our model is similar to PageRank wherein if one does a random walk over the social interaction network he would end up on important profiles more often than non important ones. We keep on iterating until the rank order does not change. In our case we observed the Imp vector to converge after 35 iterations. We used the *Jacobi* iteration method for computation of importance scores [29]. We continue our study by examining some extensions to the basic random walk approach of determining importance.

## 2. Weighted Random Walk Importance Model

The basic model of ranking profiles has some drawbacks: for instance it does not take into account the frequency of interactions. Going by the basic model a user who tweets a lots of different profiles would be considered more important than a user who converses with a few profiles but interacts with those selected profiles a lot.

In order to overcome this shortcoming we modify our basic model to take into account the edge weights of the social interaction network. In our modified model of random walk the edge weights (which signify the frequency of tweets exchanged between vertex nodes) are also incorporated in the score computation. Thus a profile will distribute its importance score to its neighbors not equally but according to the frequency with which they interact with their neighbors in the past.

The importance thus becomes:

$$Imp_u = c \sum_{v \in B(u)} Imp_v * W(u,v) \tag{4.3}$$

where *w(u,v)* represents the interaction frequency between profiles u and v. Equivalently, in matrix form:

$$Imp = \alpha.T.W_{u,v} \tag{4.4}$$

where W is the Matrix consisting weights for each u,v.

### 3.   Incorporating Trust

We also calculate user importance based on another model that has its origins in Trust rank [7] algorithm. In this model we bias our random walk on the social interaction network based on a static trust score. This profile specific trust score is assigned on the basis of some static property of the profile. One metric that we use in our computation is the number of followers of a profile. A profile having large number of followers would be more trusted in general than a profile having few followers.

User importance vector in this case would be:

$$Imp = \alpha.T.Imp + (1 - \alpha).Pref \tag{4.5}$$

where Pref is a normalized static score distribution vector of non-negative entries summing up to one. It assigns a non-zero static trust score to profiles based on their follower count. This score is then spread during the iterations to the profiles they have replied to. This biases the random walk computation in favor of profiles having high values of follower counts and helps us to combine both the *perceived* user importance of profiles among peers (like those of actors, sportsman) to the *algorithmic* importance

determined by the random walk importance model.

We also augment this basic trust based approach by incorporating edge weights into the computation. These edge weights represent the frequency of interactions between two users, or in other words an edge weight on an edge from user A to user B indicates number of times A has replied to user B.

The importance score thus becomes:

$$Imp_u = c \sum_{v \in B(u)} Imp_v * W(u,v).Pref \qquad (4.6)$$

where $w(u,v)$ represents the interaction frequency between profiles u and v and Pref is the normalized trust vector

## C. Results and Analysis of Random Walk Based Approaches

In order to find out the effectiveness of our technique we collect the top 6000 profiles retrieved by each of the proposed importance algorithms and compare them against each other. Comparing our results against the top profiles by the number of incoming tweets helps us to illustrate how our criteria for modeling user importance through a random walk is better than calculating importance by simply counting the number of incoming tweets.

While there is no standard criteria of evaluating user importance we do some manual inspection of the top profiles returned by our algorithm to get an idea of the effectiveness of our approach. We observe some very intuitive results that seem to suggest that our algorithms are effective in evaluating user importance. The results are for the Jan 31-Feb 19th dataset.

Tables X and XI shows some of the top profiles that we got after running the vanilla Random Walk model and the Weighted Random Walk model on the social

Table X. Top profiles by random walk importance

| Profile | Follower count | Rank | Description |
|---|---|---|---|
| stephenfry | 149369 | 1 | Famous writer, British actor |
| Schofe | 51051 | 2 | Philip Schofield-British television presenter |
| CHRISDJMOYLES | 61567 | 3 | Chris Moyles-English Broadcaster, author and DJ |
| Wossy | 99176 | 4 | Jonathan Ross- British television,radio presenter |
| aplusk | 63448 | 5 | Profile of Ashton Kutcher, famous hollywood actor |
| lancearmstrong | 59042 | 6 | Seven time Tour de France winner |
| richardpbacon | 9403 | 7 | BBC Radio Fivelive presenter |
| ijustine | 52966 | 8 | Avid Twitter user and blogger |
| wilw | 73775 | 9 | Will Wheaton- American actor and writer |
| scott_mills | 8959 | 10 | English radio DJ |
| levarburton | 25620 | 11 | American actor, director and educator |
| mrskutcher | 15421 | 12 | Profile of Demi Moore, famous hollywood actress |
| THE_REAL_SHAQ | 49913 | 13 | Profile of Shaquile O Neal, NBA star |

interaction network. In the table for the Random Walk model, several celebrities figure amongst the top results. These results are slightly different than the ones obtained by the top tweet count method. In addition to writers and bloggers we also observe people from other walks of life like sportsmen, actors etc. The top 13 profiles that we obtained after the Random Walk importance model are indeed very popular. Results of the Weighted Random Walk model, however, are still dominated by bloggers, and writers like the top tweet count results. This can be attributed to the frequent tweeting characteristics of these profiles. As a result of more profile activity around their profiles (more incoming tweets) and high weights on their incoming edges more importance gets transferred to them as compared to profiles who do not receive as many tweets. In the list for Weighted Random Walk importance, we also observe some profiles who have protected information about their biography because

Table XI. Top profiles by weighted random walk importance

| Profile | Follower count | Rank | Description |
|---|---|---|---|
| Sugarwilla | 507 | 1 | Small time actress |
| vojha | 372 | 2 | Blogger |
| ninjen | 970 | 3 | Profile of a blogger |
| brentschooley | 220 | 4 | iphone applications developer |
| radiojen | 39 | 5 | Appears to be a radio jockey |
| Lynnie36 | 96 | 6 | Profile no longer exists |
| Spoonsie | 90 | 7 | Blogger |
| Slugger41 | 1131 | 8 | Not enough information |
| chriswalts | 244 | 9 | Tv show host based in canada |
| TidyCat | 317 | 10 | Not enough information |
| DaveJMatthews | 28122 | 11 | African-American Grammy Award winning musician |
| tygerbaby | 1145 | 12 | Not enough information |
| RightGirl | 909 | 13 | Blogger |

of which it becomes difficult to get an accurate idea about their profile importance. Also most of these bloggers and writers are pretty regular users, while examination of their blogs do suggest a keen interest of other people in their blog posts we do not find any information in their profiles which can account for such a high rank on the Weighted PageRank results. The only possible explanation is that perhaps they interact very closely and frequently with a select few people and because of the higher edge weights they get a majority of the importance share from their neighbors.

Tables XII and XIII shows the top profiles after running the Trust-based Random Walk importance model and the Trust-based Weighted Random Walk model. The results of the Trust-based model seem to be quite similar to those of the vanilla Random Walk model with stephenfry, aplusk and wossy being the common profiles amongst the top 5 users. Results of the Trust-based model seems to capture the notion of importance correctly with a good mix of writers, bloggers and media personalities.

While the top 13 results for Weighted Trust-based random walk model are exactly identical to the Weighted random walk model results, a closer examination of the list

Table XII. Top profiles by trust-based random walk model

| Profile | Rank | Description |
|---------|------|-------------|
| stephenfry | 1 | Famous writer, British actor |
| aplusk | 2 | Profile of Ashton Kutcher, famous hollywood actor |
| Wossy | 3 | Jonathan Ross- British television,radio presenter |
| hodgman | 4 | John Hodgman- American voice actor, author and humorist |
| guykawasaki | 5 | Co-founder of social network Alltop. |
| kevinrose | 6 | Founder of social media site Digg |
| wilw | 7 | Will Wheaton- American actor and writer |
| chrisbrogan | 8 | President, New Marketing Labs(media agency). |
| Schofe | 9 | Philip Schofield-British television presenter |
| pop17 | 10 | Sarah Austin-alternative media producer and online life-caster |
| Veronica | 11 | V Belmont-co-host of the Revision3 show Tekzilla |
| lancearmstrong | 12 | Seven time Tour de France winner |

of 6000 top users show some slight differences between the two algorithm results. However the two lists are still very much identical having an overlap of about 80%. One reason for the large number of common results between weighted versions of vanilla random walk model and Weighted Trust-based random walk is that the top profiles obtained by Weighted random walk model have nearly the same number of followers. Since we biased the random walk by the number of followers of a profile ,having comparable number of followers mitigates the effect of biasing the importance score computation by a initial static trust score. The scores are still dominated by scores obtained by doing a random walk on the linked structure and are not influenced a lot by the initial trust vector biasing. However as we go down the list the difference in the number of followers of profiles leads to some profile getting a different rank in Weighted Trust-based random walk than the Weighted random walk model.

Finally, we also considered a random walk approach over the declared social network (instead of the social interaction network) to better understand if the random walk style approach would generate reasonable importance rankings. This approach however suffers from the same limitations as the naive approach based on follower

Table XIII. Top profiles by trust-based weighted random walk model

| Profile | Follower count | Rank | Description |
|---|---|---|---|
| Sugarwilla | 507 | 1 | Small time actress |
| vojha | 372 | 2 | Blogger |
| ninjen | 970 | 3 | Profile of a blogger |
| brentschooley | 220 | 4 | iphone applications developer |
| radiojen | 39 | 5 | Appears to be a radio jockey |
| Lynnie36 | 96 | 6 | Profile no longer exists |
| Spoonsie | 90 | 7 | Blogger |
| Slugger41 | 1131 | 8 | Not enough information |
| chriswalts | 244 | 9 | Tv show host based in canada |
| TidyCat | 317 | 10 | Not enough information |
| DaveJMatthews | 28122 | 11 | African-American Grammy Award winning musician |
| tygerbaby | 1145 | 12 | Not enough information |
| RightGirl | 909 | 13 | Blogger |

count, it does not take profile activity into account and is susceptible to follower-spam. These results over the traditional friendship (following relation) graph are shown in Table XIV.

The top profiles that we obtained by using the random walk model do not seem to be very important. Moreover most of the profiles in our top results are profiles of users of other countries communicating in a foreign language, because of which it becomes difficult for us to get an idea of their importance.

In the results the top profile is of kemptownben who is a city councillor of Queen Park, Brighton, UK and a parliamentary candidate. While he does appear to be a popular politician it is unclear why he is the top result. His high rank could perhaps be attributed to the fact that majority of his followers appear in our crawled dataset.

We could perhaps get a better idea of profile importance if we construct the graph over follower relations. However its impractical to do so as typically there could be millions of followers of a profile making the PageRank computation very expensive due to the huge size of the webgraph.

Table XIV. Top profiles obtained by random walk on the friend graph

| Profile | Rank | Description |
|---------|------|-------------|
| KemptownBen | 1 | Green Party activist, politician, socialist and councillor(Brighton, UK) |
| tonio888 | 2 | Not enough information from profile/regular user |
| contemplation | 3 | Regular user/Not enough information |
| jaifaitunreve | 4 | Regular user/Not enough information |
| William | 5 | Regular user |
| _misc987 | 6 | Regular user |
| xcazin | 7 | French Blogger |
| remiforall | 8 | Blogs about Web, Art, Music etc |
| Wimby | 9 | Protected information |
| fumetsuka | 10 | Protected information |
| roadtohappiness | 11 | Blogger/regular user |

D.  Rank Correlation

Finally, we investigate the relative rankings generated by different techniques over datasets of different time periods and also compare the ranks obtained over the same time period. This analysis reveals how the different methods compare, and how these methods adapt to change in frequency of interaction and profile activity. To get an idea of correlation between rankings of different algorithms we computed the Kendall tau coefficient for rankings produced by the same algorithm over different time periods and also rankings of different algorithms over the same time period. The main motivation for computing the Kendall tau coefficients was to get a sense of how similar the rank order was. In case there is high similarity we can use a combination of one or two approaches to detect user importance.

The Kendall tau coefficient is defined as:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{4.7}$$

where $n_c$ is the number of concordant pairs, and $n_d$ is the number of discordant pairs in the data set. In statistics, a concordant pair is a pair of a two-variable (bivariate)

observation data-set $X_1, Y_1$ and $X_2, Y_2$, where:

$$sgn(X_2 - X_1) \ = sgn(Y_2 - Y_1) \tag{4.8}$$

Correspondingly, a discordant pair is a pair, as defined above, where

$$sgn(X_2 - X_1) \ = -sgn(Y_2 - Y_1) \tag{4.9}$$

and the sign function, often represented as sgn, is defined as: $sgnx = -1$ if x is less than 0, 0 if x=0 and 1 if x us greater than 1.

### 1.   Comparing Importance Rankings across Different Time Periods

**Random walk Rank Correlation:** Between the datasets of Jan 31st - Feb 19th and Feb 19th- Mar 8 there were 52% common users in the top ranked users list with a Kendall tau coefficient of 0.465. Between the datasets of Feb 19th-March 8 and Mar 8- March 19th there were 47% common users in the top ranked users list with a Kendall tau coefficient of 0.435. This high overlap in the random walk model rank is consistent with our expectations. This is because the random walk model takes into account global importance (profile activity) and not just profile activity in the immediate neighborhood, even if some immediate users stop tweeting an important profile over time there will still be many incoming edges from other profiles. A high value of Kendall tau rank coefficient between the two lists suggests that the relative rank order is almost the same in both the lists showing that rank order does not change a lot in a small period.

    **Trust Based Random walk Rank correlation:** Between the datasets of Jan 31st - Feb 19th and Feb 19th- Mar 8 there were 60% common users in the top ranked users list with a Kendall tau coefficient of 0.51. Between the datasets of Feb 19th-

March 8 and Mar 8- March 19th there were 55% common users in the top ranked users list with a Kendall tau coefficient of 0.52. The high overlap in the rank lists can be attributed to the same reason as the overlap in case of basic random walk model. The percentage of overlap is higher in this case because Trust-based random walk scores are also influenced by the follower count (as we have used number of followers to bias our trust based random walk computation). Typically number of followers of important profiles grow by the similar rate and do not change a lot within 15 days.

**Top followers list correlation:** Between the datasets of Jan 31st - Feb 19th and Feb 19th- Mar 8 there were 82% common users in the top follower users list with a Kendall tau coefficient of 0.78. Between the datasets of Feb 19th-March 8 and Mar 8- March 19th there were 81% common users in the top follower users list with a Kendall tau coefficient of 0.80. The high overlap in top follower count is expected as the rate of increase of followers of important profiles remains mostly same. It might however, get influenced by some special events , like competition between profiles of cnnbrk(cnn) and aplusk(Ashton Kutcher) to reach 1 million followers. Similarly, the high values of the Kendall tau coefficients can be explained due to the fact that the relative rank order of the top followers does not change as number of followers grow by roughly the same amount.

**Top Tweet receivers(total) correlation:** Between the datasets of Jan 31st - Feb 19th and Feb 19th- Mar 8 there were 22% common users in the top tweet receivers list with a Kendall tau coefficient of 0.175. Between the datasets of Feb 19th-March 8 and Mar 8- March 19th there were 17% common users in the top tweet receivers list with a Kendall tau coefficient of 0.10. The low overlap in the top tweet receivers list can be attributed to the frequently changing interaction dynamics.The relative rank order also changes significantly due to this very reason.

**Weighted Random walk Rank Correlation:** Between the datasets of Jan

31st - Feb 19th and Feb 19th- Mar 8 there were 43% common users in the top ranked users list with a Kendall tau coefficient of 0.28. Between the datasets of Feb 19th-March 8 and Mar 8- March 19th there were 17% common users in the top ranked users list with a Kendall tau coefficient of 0.059.

The lowest overlap in the Top Tweet receivers lists can be attributed to the fact that this algorithm takes into account edge weights in the incoming replies. As edge weights are dependent on the frequency of interaction which changes a lot over brief periods of time it affects the rank computation. Sometimes a user might get involved in some conversation with his friend which might lead to high weight on the reply to edge between the users for that time period of observation, at other times he might not tweet this particular user at all or tweet him very less leading to a lower edge weight.

### 2.  Comparing Importance Rankings Generated by Different Approaches

In this section we investigate the relative rankings generated by different techniques over the same time period. This helps us to understand how differently the algorithms model the notion of importance.

Table XV shows the percentage of overlap between the top 6000 users obtained by different algorithms. After analyzing this table we observe that there is a high amount of overlap between Weighted random walk rank and Weighted Trust based random walk model. This can be attributed to the fact that most of the top profiles that figure high on the Weighted random walk list have comparable number of followers. The effect of biasing rank computation thus does not have much effect on the ranking.

Out of these lists top tweet frequency count based method has the least number of common profiles to other lists. This can be attributed to the fact that it is just based on profile activity which cannot necessarily capture importance that effectively,

Table XV. Percentage of overlap in user ranks between different algorithms

| Algorithm | Weighted random walk | Basic random walk | Trust-based random walk | Weighted Trust based random walk | Top Followers | Top Tweet Recievers |
|---|---|---|---|---|---|---|
| Weighted random walk | 100 | 19 | 19 | 83 | 10 | 6 |
| Basic random walk | 19 | 100 | 55 | 22 | 28 | 6 |
| Trust-based random walk | 19 | 55 | 100 | 23 | 49 | 7 |
| Weighted Trust based random walk | 83 | 22 | 23 | 100 | 14 | 6 |
| Top followers | 10 | 28 | 49 | 14 | 100 | 7 |
| Top Tweet Receivers | 6 | 6 | 7 | 6 | 77 | 100 |

as discussed earlier in our results. The high overlap between top follower list and top tweet receiver list can be attributed to the fact that profile having large number of followers get more incoming reply tweets as their status/ tweets reach a large audience due to the high number of followers subscribing to them.

On the other hand random walk models all utilize the same reply graph and due to the nature of these algorithms a profile that is very popular will get a mention in almost all the different flavors of the random walk. To get an idea of how these algorithms rank profiles differently we compared the ranks of some preselected profiles obtained by different algorithms which is shown in Table XVI N/A here stands for the case when the profile does not figure in the top 6000 list. stephenfry is ranked first in almost all of the algorithms because he has large number of followers as well lots of incoming messages, his blog posts, status messages attract a lot of replies and interest amongst his followers.

Table XVII shows the Kendall tau correlation coefficient between the different

Table XVI. Ranks of common profiles obtained by different algorithms

| Profile | Nature of Profile | Weighted random walk | Basic random walk | Trust based random walk | Top Followers | Top Tweet Recievers(Total frequency) |
|---------|-------------------|----------------------|-------------------|-------------------------|---------------|--------------------------------------|
| cnnbrk | Major News channel | N/A | 1019 | 18 | 3 | N/A |
| aplusk | Hollywood actor | 424 | 5 | 2 | 18 | N/A |
| masterconsole | Spam Profile | N/A | N/A | N/A | N/A | 1 |
| lancearmstrong | Famous sporstman | 1639 | 6 | 12 | 21 | N/A |
| kevinrose | CEO of Digg | 1718 | 17 | 6 | 5 | N/A |
| stephenfry | Famous Blogger | 720 | 1 | 1 | 1 | N/A |

rankings by algorithms over the same time period.

From this table we observe that the highest Kendall tau coefficient was between the Weighted Trust based random walk and Weighted random walk results. This shows that the two lists agree highly on the rank order. The amount of overlap in the lists is also high. Both these factors suggest that the two methods are very similar. The reason for the high overlap and high value of Kendall tau coefficient can be attributed to the fact that the top profiles in case of Weighted random walk have comparable number of followers because of which if we bias our random walk by the number of followers(as in Weighted Trust based random walk) it does not lead to a significant change in the rank order or the top profiles.

The high Kendall tau coefficient between the top Followers list and the top Trust based random walk results can be attributed to using follower count to bias the random walk computation in case of Trust based random walk.

Table XVII. Kendall Tau coefficients for different algorithms over same period

| Algorithm | Weighted random walk | Basic random walk | Trust based random walk | Weighted Trust based random walk | Top Followers | Top Tweet Receivers |
|---|---|---|---|---|---|---|
| Weighted random walk | | 0.01 | 0.06 | 0.74 | 0.11 | -0.04 |
| Basic random walk | 0.01 | | 0.35 | 0.21 | 0.26 | 0.397 |
| Trust based random walk | 0.06 | 0.35 | | 0.032 | 0.41 | 0.146 |
| Weighted Trust based random walk | 0.74 | 0.21 | 0.032 | | 0.082 | -0.14 |
| Top followers | 0.11 | 0.26 | 0.41 | 0.08 | | -0.14 |
| Top Tweet Receivers | -0.04 | 0.11 | 0.146 | -0.014 | -0.14 | |

We also observe that mostly negative values of Kendall tau coefficient are observed between top Tweet receivers list and other algorithms. This coupled with the fact that we see least overlap in the top Tweet receivers list and other algorithms suggests that this method is the least similar to other methods. This is a further proof of the fact that this method does not capture importance effectively as it is based on the profile activity alone and is thus highly prone to spam, and bots. The ranking generated by this algorithm has the least agreement with other techniques.

In this section we discussed various approaches to model user importance based on the interaction graph along with giving their mathematical background. We compared results of the various algorithms to get an idea of how differently each algorithm models importance. Our results suggest that out of the various approaches a random walk based approach incorporating trust is closest to our notion of user importance

as it takes into account both interaction characteristics of users along with incorporating the notion of trust. It can also be modified to incorporate different flavors of importance. For example in our calculation we biased the random walk by the number of followers, if we want to get users that are more active on twitter we can bias the computation by the number of incoming tweets. We can also assign initial trust scores based on other criteria like number of profile views if the social network in consideration has support for finding that out. This method can thus be used to find importance according to the desired property, it incorporates the advantages of both the random walk models and the static naive approaches.

In the following section we discuss some possible extensions of our work.

CHAPTER V

CONCLUSIONS AND FUTURE WORK

In this thesis we have presented the first large-scale study of the dynamic interactions on a real-world Social Information System. Our study over the Twitter microblogging service has examined the usage statistics, growth patterns, and user interaction behavior of over 2 million participants. Based on the analysis of the interaction structure of Twitter, we have explored several approaches for measuring user importance by leveraging the interaction dynamics of the social network. We have seen how interaction-based user importance measures differ from traditional static network measures.

Determining user importance is especially important on the emerging social computing framework, as users are now central to content creation, annotation, and rating. Based on our observations in this thesis, we conclude by presenting a roadmap for future studies of how user importance can be incorporated into applications centered on Social Information Systems. We consider three potential applications of our research: in detecting spam [22], in building more sophisticated recommendation and advertising systems [24], and in improving the quality of social event detection.

**Spam detection:** Our study on detecting user importance can help in detecting spam profiles on Twitter. Spam profiles in Twitter are typically those that follow a large number of other profiles and send unwanted tweets. *Follower* spam is also getting to be quite common in Twitter. In this type of spam the spammers send a request to large number of profiles asking for permission to follow them, out of these profiles a small percentage of profiles like to reciprocate this request and follow these spammers back, not realizing that these are spam profiles. They would then see all the spam messages on the spammer's profile. Sometimes spammers prefix the tweets

by @username, which then causes the tweets to show up in the timeline of the profile of the user. There have been recent attempts by Twitter to address this issue by adding a verified attribute to a profile which shows up on the homepage of the profile letting users know that the profile is a genuine one. However this is not enough.

By constructing the social interaction network, spam profiles can be identified; typically a spam profile would have a large outdegree in the social interaction network but a low algorithmic importance. For example, in the Twitter-based interaction network described in Chapter III, the node having the maximum outdegree MStweet, happens to be a spam profile. The underlying link structure of the social interaction network coupled with the fact that it reflects the current snapshot of interactions between profiles can be effectively used in spam detection. Another way of detecting spam would be to construct the transpose of the social interaction network graph and run PageRank. The high ranked profiles which follow a large number of other profiles would be then ideal candidates of spam profiles.

**Advertising and recommendation systems:** In the context of the immense advertising potential of Twitter, it is important to gather information about profile activity and its importance among its peers. Another reason why Twitter and in general other microblogging networks are so critical to advertising is that the very nature of information exchange on this network has a very real-time character to it. Users usually tweet about what they are doing or about some products they find interesting. Advertisers could use this information to recommend their products to those users. If we had a mechanism that could broadcast advertisements from the so called "important" profiles it can have a huge potential to increase advertising revenue. People are more likely to follow recommendation of profiles they consider important and take their advice about a company product than a direct advertising campaign coming from a company.

**Event Detection and community detection:** Profile activity of important profiles, like the topic that profile tweets about, can also help in determining importance of events. An event talked about by majority of important profiles can be safely assumed to be important to majority of people. This is especially relevant to the conversational nature of Twitter. Since the social interaction graph consists of several strongly connected components, it would be interesting to find out what all those profiles in the strongly connected component tweet about, and how these communities grow.

REFERENCES

[1] J. Callan, J. Allan, C. Clarke, S. Dumais, D. Evans, M. Sanderson and C. Zhai,"Meeting of the minds: An information retrieval research agenda," *SIGIR Forum*, 2007, pp. 25-34.

[2] S.A. Golder, D. Wilkinson and B.A. Huberman, "Rhythms of social interaction: Messaging within a massive online network," *Third International Conference on Communities and Technologies*, at http://www.hpl.hp.com/research/idl/papers/facebook/facebook.pdf, 21 December 2008.

[3] B. Krishnamurthy, P. Gill, M. Arlitt, "A few chirps about Twitter," *in Proc. of the First Workshop on Online Social Networks*, Seattle WA, August 2008, pp. 19-24.

[4] B. Huberman, D. Romero, F. Wu, "Social Networks that Matter: Twitter under the Microscope," *First [Online]*, vol. 14, No. 1, January 5, 2009.

[5] A. Java, X. Song, T. Finin, B. Tseng, "Why we twitter: Understanding microblogging usage and communities," *in Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis.*, August 2007, San Jose, CA, pp. 56-65.

[6] E. Garcia, M. Silica, T. Calvo, "Evaluating web document quality with linguistic variables: Combining informative and page design quality," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Italy, July 2004, pp. 1975-1982.

[7] Z. Gyongyi, H. Garcia-Molina, J. Pedersen, "Combating web spam with TrustRank," *30th International Conference on Very Large Data Bases* (VLDB 2004), Toronto, Canada, August 29 - September 3, 2004, pp. 576 - 587.

[8] L. Page, S. Brin, R. Motwani, T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *Technical Report*. Stanford InfoLab, Palo Alto,CA, 1999.

[9] W. Xing, A. Ghorbani, "Weighted PageRank algorithm," *Second Annual Conference on Communication Networks and Services Research* (CNSR'04),Fredericton, N.B.,Canada, 2004, pp. 305-314.

[10] A. Broder, R. Kumar, F. Maghoul, P. Raghavan,S. Rajagopalan,R. Stata, A. Tomkins and J. Wiener, "Graph structure in the web," *Computer Networks-The International Journal of Computer and Telecommunications Networking, vol. 33, Issue 1-6* (June 2000),Holland, pp. 309 - 320.

[11] H. Chen, "Homeland security data mining using social network analysis," *in Proc. of the 1st European Conference on Intelligence and Security Informatics*,Denmark, 2008, pp. 4–4.

[12] L. Grossman, "Iran protests: Twitter, the medium of the movement," *Time Magazine*, Jun 17, 2009.

[13] C-F. Hsu, E. Khabiri, J. Caverlee, "Ranking comments on the social web," *The 2009 International Conference on Social Computing (SocialCom-09)*, Vancouver, Canada 2009, pp. 27-35.

[14] Twitter. [Online]. Available: http://twitter.com/public_timeline.

[15] J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM (JACM)*, archive vol. 46, No. 5, September 1999, pp. 604 - 632.

[16] E. Agichtein, C. Castillo, D. Donato, A. Gionis, G. Mishne, "Finding high-quality content in social media," *in Proc. of the International Conference on Web Search and Web Data Mining*, 2008, Palo Alto, CA, pp. 183-194.

[17] J. Zhang, M.S. Ackerman, L. Adamic, "Expertise networks in online communities: Structure and algorithms," *in Proc. of the 16th International Conference on World Wide Web*, Banff, Alberta, Canada, 2007, pp. 221 - 230.

[18] C. S. Campbell, P. Maglio, A. Cozzi, B. Dom, "Expertise identification using email communications," *in Proc. of the 12th International Conference on Information and Knowledge Management* , New Orleans, LA, 2003, pp. 528 - 531.

[19] R. Kumar, J. Novak, A. Tomkins, "Structure and evolution of online social networks," *in Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, 2006, pp. 611 - 617.

[20] B. Dom, I. Eiron, A. Cozzi, Y. Zhang, "Graph-based ranking algorithms for e-mail expertise analysis," *in Proc. of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, San Diego, CA, 2003, pp. 42 - 48.

[21] R. Guha, R. Kumar, P. Raghavan, A. Tomkins, "Propagation of trust and distrust," *in Proc. of the 13th International Conference on World Wide Web*, New York, NY, 2004, pp. 403 - 412.

[22] M. Glaser, "How follower spam infiltrated twitter and how to stop it," [Online]. Available: http://www.pbs.org/mediashift/2008/10/how-follower-spam-

infiltrated-twitter-and-how-to-stop-it297.html, *Mediashift*, October 2008.

[23] Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Twitter.

[24] K. Shinkle, "Why twitter advertising could be a huge success," [Online]. Available: http://www.usnews.com/blogs/the-ticker/2009/04/21/why-twitter-advertising-could-be-a-huge-success.html, *US News and World Report*, April 2009.

[25] L. D'Monte, "Swine flu's tweet tweet causes online flutter," [Online]. Available: http://www.business-standard.com/india/news/swine-flu

[26] "Twitter.com - Traffic details from Alexa," [Online]. Available: http://www.alexa.com/siteinfo/twitter.com, *Alexa Internet*, July 2009.

[27] R. Kelly, "Twitter Study - August 2009," Twitter Study Reveals Interesting Results About Usage, San Antonio, Texas: *Pear Analytics*, August 2009.

[28] Quantcast. [Online]. Available: http://www.quantcast.com/twitter.com, October 2009.

[29] Suchmaschinen Doktor. [Online]. Available: http://pagerank.suchmaschinen-doktor.de/matrix-inversion.html, November 2009.

[30] J. Flintoff, "Thinking is so over," [Online]. Available: http://technology.timesonline.co.uk/tol/news/tech_and_web/personal_tech/article1874668.ece, *Times Online*, June 2007.

VITA

Anupam Aggarwal received his Bachelor of Technology (B.Tech) from the National Institute of Technology, Kurukshetra, India in 2006. He worked for a communications software company, Aricent Inc, India from 2006-2007 before finally coming to Texas A&M to pursue his Master's degree. He is a founding member of the TAMU infolab focused on Web and Distributed Information Management. Anupam's research interests include information retrieval, large-scale systems, pattern classification and the Web. He graduated from Texas A&M in December 2009.

He may be contacted at the following address: Department of Computer Science, 301 HRBB, Texas A&M University, College Station, TX 77843-3128.

The typist for this thesis was Anupam Aggarwal.