Social Status and Role Analysis of Palin's Email Network

Xia Hu
Arizona State University
Tempe, AZ 85287, USA
xiahu@asu.edu

Huan Liu
Arizona State University
Tempe, AZ 85287, USA
huanliu@asu.edu

ABSTRACT

Email usage is pervasive among people from different backgrounds, and can be an important and accurate data source to study intricate social structures. Social status and role analysis on a personal email network can help reveal hidden information. The availability of Sarah Palin's email corpus presents a great opportunity to study the social statuses and social roles in an email network. However, the email corpus does not readily lend itself to social network analysis due to problems such as noisy email data, scale in size, and temporal constraints. In this paper, we contribute an initial investigation of social status and role analysis on Sarah Palin's email corpus. In particular, we reconstruct a multiplex network from the unstructured email corpus, and then analyze the social statuses and roles from three different perspectives: individual, group, and temporal. Experimental result demonstrates that our proposed analytic tool provides an effective way to analyze social status and roles on email networks.

To the best of our knowledge, this work is the first quantitative study of Sarah Palin's email corpus recently released by the state of Alaska.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Science**]: Economics, Sociology; D.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Human Factors, Measurement

Keywords

Email Corpus, Social Status, Social Role Analysis, Evolution

1. INTRODUCTION

Email communication is an early form of social media. Email usage is pervasive among people of different ages and backgrounds, and is among the most common activities on the Internet. A Pew Internet and American Life report[1] in December 2010 showed that 92% of American adult Internet users send or read emails online, making email communication the most common way to communicate among all kinds of social media. The personal email network is an important and accurate source to reflect social relationships of an individual [23].

Social status is the degree of honor or prestige attached to one's position in society [9]. Most societies have some form of social hierarchy with some people in stronger, more dominant positions, and others in weaker, lower positions. Sociologists have discussed the issues to determine the positions of individuals, with their social roles, occupy in the status structure of a given society [10]. Studying social status and the corresponding social roles in a personal email network is helpful for people to manage their social resources. For example, by clearly understanding the social statuses and roles in our own email network, we can adjust our biases such as ignoring some important individuals, groups or social ties in our social network. However, real-world personal email data is rarely publicly accessible due to privacy issues, making social status and role analysis in email networks still a fertile field to be explored [8].

Recently, the state of Alaska released thousands of emails sent and received by former governor Sarah Palin. This email corpus consists of 24,199 printed pages of emails (totaling 250 pounds of paper) covering Palin's first 22 months as governor – from December 2006 until she accepted the vice-presidential nomination in August 2008. The corpus is appealing to the public and mainstream media, including CNN, NYTimes, BBC, etc. because it is 1) a large scale 2) personal email collection 3) from a politician 4) over a period of two years. The mainstream media requested the public to review the documents, saying "We invite our readers to examine them and contribute to the discussion[2]" by CNN and "Help Us Review the Sarah Palin E-Mail Records[3]" by NYTimes.

CNN and other mass media asked people to examine the data, but they failed to provide any tools to deal with such a big corpus. Although people can review the documents, it is difficult for the public to go over thousands of documents, in which many of them are intrinsically connected, and the connections may change over time. First, the large-scale networks with many nodes and complex interactions inhibit efficient processing of the network. Second, the network is built over two years and many events have occurred during that period. Also, it is imprudent to examine the network

[1]http://www.pewinternet.org/trend-data/online-activites-total.aspx

[2]http://www.cnn.com/2011/POLITICS/06/10/alaska.palin.emails

[3]http://thecaucus.blogs.nytimes.com/2011/06/09/help-us-investigate-the-sarah-palin-e-mail-records/

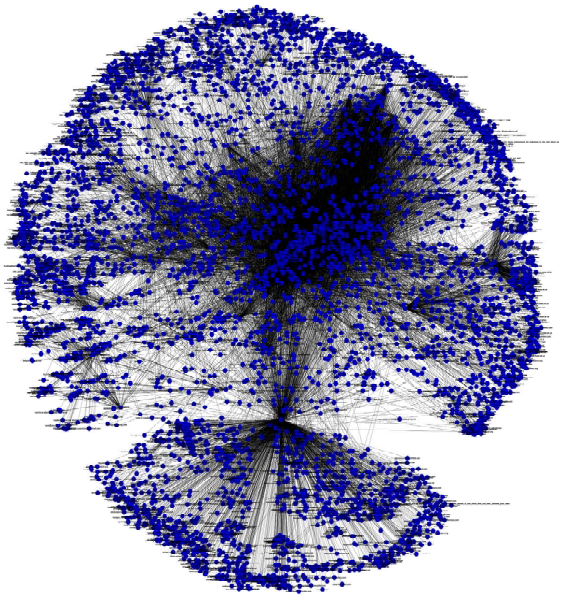


Figure 1: Visualization of Palin's Email Network

directly due to the temporal constraints. In Figure 1, we depict the email network based on a sending/receiving relationship. Although we can see two groups clearly from the global visualization, it is still insufficient to understand the email network without a finer level analysis. The Christian Science Monitor has doubted if we can successfully mine valuable information from the corpus "Sarah Palin emails: Treasure trove or waste of paper?[4]". In many aspects, it is appealing for our community to provide an effective analytic tool to help analyze Palin's email network.

In this paper, we address a novel problem to understand the social statuses and social roles of people in Sarah Palin's email network. We provide an analytic tool that enables a person/analyst to possibly study a massive number of emails, discover and study different questions in social media, that would otherwise be impossible or difficult. Our study starts by reconstructing a multiplex network from the noisy and unstructured email corpus. Then, the reconstructed network is investigated from individual, group, and temporal perspectives. We aim to understand the social statuses and social roles in Palin's email network by answering following questions.

- **Individual analysis:** What is the social status of Palin? Who are the individuals with high social status? What are the social roles of these individuals?
- **Group analysis:** Who are the individuals with high social status in the groups? Do the groups have social statuses as well?
- **Temporal analysis:** Do the key individuals in the network always have high social statuses? How do the social statuses of the key individuals evolve along the timeline? Do they change significantly during a specific time period?

[4]http://www.csmonitor.com/USA/Politics/2011/0611/Sarah-Palin-emails-Treasure-trove-or-waste-of-paper

This paper is organized as follows. Section 2 introduces data preparation and network reconstruction based on the original email corpus. In Section 3, we analyze Sarah Palin's email network from three different perspectives to understand the social statuses and roles in Sarah Palin's email network. We briefly review the related work in Section 4 and conclude in Section 5. Key individuals and their positions are indexed in the Appendix.

2. DATA PREPARATION

The Palin documents contain aspects of Sarah Palin's governorship of Alaska. They were made public by the Alaska governor's office on June 11, 2011. A scanned copy of the 24,000 pages of emails were then made available by CNN[5]. To facilitate the study of the corpus, we build an XML corpus based on the public dataset. In this section, we describe how to build the XML corpus and reconstruct a multiplex network from the corpus.

2.1 XML Dataset Preparation

The Alaska governor's office only provided an unstructured hard copy of the emails to the public. We first refine the raw corpus. As the emails are already scanned based on the hard copy as a set of images, we employ a piece of Optical Character Recognition (OCR) software provided by Adobe Acrobat to convert all images to text documents.

An email has several fields, such as title, sender, receiver etc. Manco et al. [17] introduced three types of features that need to be considered in email related tasks: unstructured text, categorical text, and numeric data. To uncover important information in the corpus, we extend the features to four categories, shown as follows:

- Unstructured text: "Body" and "Subject" of the email;
- Categorical text: email creation time ("Sent");
- Numeric data: message size, number of recipients;
- Relationship data: sender and receiver of the email ("From", "To", "CC").

To make the data easily accessible, we build an XML corpus[6] based on the raw text corpus. Each email (E) is indexed as: *E = <"Body", "Subject", "Sent", "From", "To", "CC", "Size", "numRecipients">*, and there may be multiple entities in "To" and "CC" fields.

In the XML corpus, there are a total of 13,170 messages belonging to 746 distinct senders and 5,232 distinct receivers. Figure 2 shows the number of emails sent by anyone in the corpus over months. The highest bar was in April 2008, the month in which she gave birth to her fifth child, Trig. Also, during January 2008 to August 2008, there were key events, including her daughter's pregnancy and marriage, the announcement of becoming John McCain's running mate, etc. Thus her email communication is very active during that period.

2.2 Name Entity Resolution

In the corpus, the quality of generated name phrases from images by OCR is low. In addition, a person intentionally

[5]http://www.cnn.com/specials/2011/palin.emails/index.html
[6]The dataset will be made available upon request for research purposes with the analytic tool.

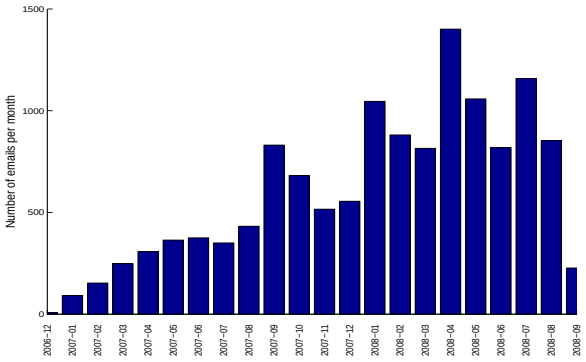


Figure 2: Numbers of Emails along timeline of the Email Corpus

or by chance uses different names in different emails. When we consider these name phrases as distinct individuals, it leads to inaccuracies for the constructed social network. We utilize textual features to address this *Name Entity Resolution (NER)* problem. The textual features, named as surface information, are widely used in similarity measurement methods [5]. It is based on the hypothesis that when two phrases have more words that overlap, they are considered to be related more. The NER problem is beyond the scope of this study. We can use any other feasible NER solution in this preprocessing step to the same effectiveness.

As the names used in "Sent" and signature fields are always formal and complete, they are extracted as the standard entity set in our study. We map all the name phrases into the standard set according to surface similarity.

Let S denote the standard set where $S = \{s_1, s_2, ..., s_m\}$ and s_i denotes a distinct name entity. Given a name phrase r_i, we compute the semantic similarity between r_i and s_j by counting the co-occurrence of words. The total occurrences of words from s_j in phrase r_i are denoted as $f(r_i \mid s_j)$; and we define $f(s_j \mid r_i)$ in a similar manner. The total number of words in r_i is denoted as $C(r_i)$, and similarly for $C(s_j)$. To calculate the surface similarity between them, a variant of a popular similarity metric [4] – Jaccard coefficient, is used as below:

$$SurfSimi(r_i, s_j) = \frac{min(f(r_i \mid s_j), f(s_j \mid r_i))}{C(r_i) + C(s_j) - max(f(r_i \mid s_j), f(s_j \mid r_i))}. \quad (1)$$

To avoid bias, we normalize all m scores using a linear normalization formula, and map our evaluated name phrase r_i into $s^\star$, which has the highest surface similarity score and is larger than a threshold θ, and is defined as:

$$s^* = \arg\max_{s_j \in S} SurfSimi(r_i, s_j), \quad (2)$$

$$s^* > \theta, \quad (3)$$

where θ controls the number of receivers needed to be mapped. In our experiment, we empirically set $\theta = 0.66$. Thus, the similar name phrases are mapped into distinct name entities.

2.3 Email Network Reconstruction

Table 1: Comparison of Networks With & Without Sarah Palin

	With Palin	*Without Palin*
# of Nodes	4446	4445
Biggest Component	4446	3773
# of Edges	59589	22177
# of Distinct Edges	13888	10021
Clustering Coefficient	0.146	0.171
Network Centralization	0.220	0.072

In the language of email communication network, vertices correspond to individuals that sent or received email, and edges correspond to sending/receiving relations. We reconstruct an email network based on <"From", "To", "CC"> fields in our XML corpus. The construction of the communication network is straightforward. The fields are parsed to extract a sender and multiple receivers. The names phrases extracted from the emails are processed using our proposed name entity resolution methods. The identified distinct entities are employed as vertices of the network. Edges are added between each pair of entities (sender and receiver). The graph edges are directed from the sender to the receiver of the email. The tie strength (link weight) between two nodes is represented as the number of emails sent and received between them. Two nodes become "closer friends" in social network when they communicate more with each other [31]. Thus, we reconstruct a weighted and directed communication network based on the email corpus.

3. STATUS & ROLE ANALYSIS

In this section, we investigate social statuses and social roles of the reconstructed email network. In Table 1, we summarize the statistical properties of the network, which is a weighted and directed graph with 4,446 nodes and 13,888 distinct edges. Clustering coefficient is a measure of the likelihood that two associates of a node are mutually connected [28]. The clustering coefficient of 0.146 indicates this network is sparse in structure. Network centralization is the difference between the number of links for each node divided by maximum possible sum of differences [28]. The network centralization of 0.220 represents this network is not strictly centered by one node as a star structure.

3.1 Individual Analysis

An important way to understand the social status structure of a social network is to study the people, especially the important people in the social network. We define the people with high social status as *key individuals* in the social network. In this subsection, we aim to analyze the characteristics of key individuals in the email networks.

3.1.1 Social Status of Palin

Obviously, Palin has the highest social status in the network. To further explore her importance to the social network, we employ a "*knockout*" technique in the experiment. Knockout based methods have been widely used in many areas, like gene function analysis [7], to test the overall performance variance brought by one process or one component when it is made inoperative in the framework. We conduct

Table 2: The Individuals with Highest In-degree and Out-degree

Out-degree (sent)		In-degree (received)	
number	name	number	name
1129	John Katz	4956	Tibbles Michael
955	Myrna Brown	4932	Janice Mason
880	Janice Mason	4332	Kristina Perry
794	Ivy Frye	3630	Leighow Sharon
713	Meghan Stapleton	1835	Nizich Michael

experiments to compare the differences brought by "knocking out" Palin from the network.

Palin and all related links are removed from the network, and we compare the statistical properties between the networks with and without Palin. As shown in Table 1, even that the total number of links in the network decreases significantly from 59,589 to 22,177, most of the nodes (84.9%) and unique links (72.2%) still exist in the biggest component. The clustering coefficient even increases without Palin, which demonstrates that the biggest component has a more robust structure when comparing to the original network. Lower network centralization means center of the network becomes more sparse. The removal of Sarah Palin will not bring in devastating effects to the email network.

3.1.2 Social Status of Individuals

Social status is recognized according to what a particular society or culture deems valuable. Indeed, various features can be employed in determining one's social status and the selection of features can be subjective. The most common and important relationship in email networks is the sending/receiving relationship. We start our discussion from the simplest case, approximating key individuals with high social status as people who send/receive emails frequently.

Once individual A sends an email to individual B, the sending/receiving relationship is built. For the ease of presentation, in this case, we define A is B's sender and B is A's receiver. An individual's out-degree and in-degree are formally defined below.

DEFINITION 1: The out-degree *outDeg(u_i)* of an individual u_i is defined as the number of distinct individuals she sent emails to.

$$outDeg(u_i) = \#\{u|(u_i \rightarrow u), u \in U\}$$

DEFINITION 2: The in-degree *inDeg(u_i)* of an individual u_i is defined as the number of distinct individuals she received emails from.

$$inDeg(u_i) = \#\{u|(u \rightarrow u_i), u \in U\}$$

Table 2 lists the top 5 individuals according to out-degree and in-degree respectively. "John Katz", the head of Palin's Washington office, sent most emails out[7]; "Tibbles Michael", the Chief of Staff in Alaska, received most emails from others. Most people in the list also play important roles in Palin's social network, like the directors of departments and important campaign aides.

[7]For the ease of presentation, we index all the individuals mentioned in this paper and their positions in Appendix.

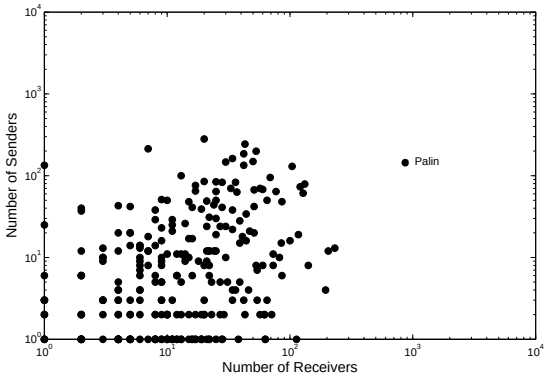


Figure 3: Number of Receivers vs. Number of Senders for Each Person in the Email Network

Table 3: Top 5 Active Individuals and Key Individuals

Active Individuals	Key Individuals
Janice Mason	Leighow Sharon
Beth Leschper	Mintz Tanci
Kristina Perry	Myrna Brown
Joseph Balash	Nizich Michael
Leighow Sharon	Meghan Stapleton

From the table, we can see there is only one overlap (Janice Mason) between top 5 individuals. There is only 22% overlap among all the top 100 individuals in the two categories. Especially, based on the top ranking individuals in the two lists, many people with high out-degree have very low in-degree and vice versa. We examine the correlation between number of receivers and number of senders for each person in Figure 3. There is a low correlation between number of senders and number of receivers, showing that the linear dependence between these two features is very weak.

We employ a linear function with equal weights to combine the normalized score of in-degree and out-degree to measure the activity of an individual. The top 5 individuals are considered *active individuals* and listed in left column of Table 3. To better utilize link structure to evaluate the social statuses of individuals, a natural way is to use PageRank [20]. Here, we define the *key individuals* as the people with high social status in the email network. The top 5 key individuals are listed in right column of Table 3. We also applied *HITS* [12] and Influence Maximization [11] to select top k individuals, and the results are very similar as those using PageRank. It is mainly because the link information plays dominant role in the email network.

Based on the observation about important individuals in the email network, they have the following patterns.

- **Hub** Individuals with high hub scores represent people that send emails to many other people. They are in charge of spreading information to others for Palin, like John Katz.
- **Authority** Individuals with high authority scores are always informed and seldom reply back, like Leighow

Table 4: Characteristics of the Two Groups With & Without of Sarah Palin

	With Palin		*Without Palin*	
	G1	**G2**	**G1**	**G2**
# of Nodes	3364	1082	3363	1081
# of Edges	32334	6255	15007	882
# of Distinct Edges	8115	5773	5886	181
Clustering Coefficient	0.105	0.098	0.096	0.018
Net. Centralization	0.448	0.778	0.301	0.067

Sharon, a spokeswoman for Palin.

- **Active** These are individuals with balanced in-degree and out-degree scores, and communicate actively. They often overlap with hub and authority individuals.

- **Key** These people are considered to be individuals with high social statuses, considering the overall network structure and unbalanced sending/receiving relationship.

We notice that key individuals do not have many overlaps with active individuals in the table. Based on the observations, we can conclude that key individuals are not necessarily equal to active individuals in the email network.

3.2 Group Analysis

In order to further understand the social statuses and social roles in the email network, we decompose this network into groups.

In this email network, as the group (a.k.a. community) membership information is not explicitly given, we need to first identify groups in the network. Since it is difficult to determine the number of groups [15] to be explored, we employ a multi-resolution community detection approach [25], instead of tuning a proper parameter. Given a network, it is zoomed into multiple levels of groups by conducting agglomerative hierarchical clustering. In this work, we focus on two levels: one has two groups and another with ten groups.

3.2.1 Contrast Analysis Between Two Groups

We start investigation from the top level of the hierarchy generated by the multi-resolution community detection method, and it consists of two groups. We summarize statistical properties of the two groups in the first two columns of Table 4: "G1" represents Group 1 and "G2" means Group 2. There are more people and more activities (number of edges) in Group 1. The clustering coefficient of Group 1 is higher than that of Group 2, indicating that the individuals in Group 1 connect well with each other. Network centralization of the Group 1 is lower, meaning there is little variation between the number of links each node possesses. Thus the network structure is more robust in Group 1 according to its lower network centralization. Since communication in the group is not centered around a very small number of nodes, the network appears more robust upon the removal of some key individuals. In addition, we calculate the key individuals in each group separately. Most of the top 200 key individuals (84%) in Group 1 are active or key individuals in the whole network. However, few people from the second group (10%) are active or key individuals in the whole network. It shows that groups also have their corresponding social statuses in the network, and the social status of Group 1 is higher than Group 2 in our case.

Table 5: Top 5 Interactions

interactions	sender and receiver
187	Janice Mason –> Kristina Perry
127	Meghan Stapleton –> Tibbles Michael
123	Janice Mason –> Tibbles Michael
103	John Katz –> Tibbles Michael
87	Meghan Stapleton –> Leighow Sharon

Table 6: Selected Groups with Clear Social Roles

	Active Individuals
Group 1_1	Todd Palin, Bristol Palin
Group 1_2	Kevin Harper, Ausman Earle
Group 1_3	Martha Rutherford, Anders Bruce
Group 1_4 *	Kim Anna, Leighow Sharon
Group 1_5	Kristina Perry, Beth Leschper

The two groups have their distinct characteristics and social statuses. Here, we aim to study the intra- and inter-group interactions. The top 5 interactions in the whole network are listed in Table 5. We note that most of the senders and receivers are from the first group, indicating that the communications in the first group are more active than those in the second. In addition, we examine the interactions between two groups and find that only 4.3% of interactions are from inter-group communication. Among inter-group communication, many (78%) are via Sarah Palin.

Based on the observations, we can draw the following conclusions. (1) Members in Group 1 are more active than those in Group 2, in both internal and external activities, further demonstrating that the social status of Group 1 is higher than Group 2. (2) Communication between the two groups are not active. (3) These two groups are basically connected because of Sarah Palin and she also has the highest betweenness centrality score.

We further study the network and groups indirectly by conducting "Knockout test" on the two groups, as shown in the right two columns of Table 4. The clustering coefficient decrease from 0.105 to 0.096 in Group 1, from 0.098 to 0.018 in Group 2. It reveals that social structure of Group 1 remains similar, and the communications in Group 2 have been significantly obstructed because of Sarah Palin's removal. The network centralization of Group 2 decreases significantly as well. It is because that the internal communications between group members are not active but only connect directly with Palin herself (with a high network centralization).

3.2.2 Key Individuals and Key Groups

We conduct experiments with a finer level of the hierarchy of groups, which consists of ten groups. Among the ten groups, eight are generated from Group 1 and two are from Group 2. In many aspects, the social role of a group can be represented by a small set of key individuals in the group. We study the social roles of each group by analyzing key individuals in the ten groups.

Key individuals are selected based on the results from

PageRank as calculated in Section 3.1. By examining the key individuals in each group, we can observe that several groups show clear social roles. For example, five representative groups generated from Group 1 are listed in Table 6. From the active individuals in the groups, Group 1_1 contains most of Palin's family members, like Todd Palin, Bristol Palin and her personal assistants. Group 1_2 contains members from various of external companies and organizations, like Kevin Haper. Group 1_3 has many members from the Department of Transportation and Natural Resources, like Martha Rutherford and Anders Bruce. Also, some groups have multiple social roles and need to be clustered into smaller groups. For example, Sarah Palin's family members (e.g. Todd Palin, Track Palin and Bristol Palin) should be clustered into the same group, but not mixed with her personal assistants, who were involved in her family arrangements (like travel etc.). We observe that members of two sub-groups generated from Group 2 does not show clear social roles.

From the ten groups, we find that a small group (with fewer members) does not necessarily have low social status in the network. Group 1_4 (with star in Table 6), which contains only 1.1% of all members (51/4446) in the whole network, has communications with different people (48.8%). Actually, many of the members from this group are recognized as key individuals in Section 3.1. This kind of group can be considered as a *key group* in the social network.

As this small group of members can connect to 48.8% people in the social network, these key individuals can be considered as the alternatives or representatives for Palin in the whole network. The reason the removal of Palin does not significantly destroy the structure of the social network is that the key individuals in the network can take charge of the information diffusion instead of her.

3.3 Temporal Analysis

As this network spans over two years, the temporal features could provide further valuable information. In this subsection, we will discuss temporal analysis of the email network.

3.3.1 Temporal Patterns of Key Individuals

The time window is fixed to three months and we take snapshots of the network with and without Sarah Palin at eight different points in time over the two years[8]. The top 5 key individuals are extracted from each network built at different time points. As some individuals have high social status along the timeline, there are eleven individuals in total who ever rank top 5 in a specific time point. The social status evolution of the eleven individuals is shown in Figure 4. The left eight columns indicate the results with Palin in the network and the right ones are the results after the removal of Palin. Each row represents a key individual's status evolution over time. The (i, j)-cell in the matrix stores the social status of the i^{th} individual in the j^{th} time window. For example, the first cell (01-2007, Myrna Brown) shows that Myrna Brown was ranked 4th at 01-2007 in the network with Palin. Blank cells with no numbers mean the individual is not ranked among the top 5 for that time period. The color gradient represents an individual's social status, with a darker color indicating a higher status.

[8]As there are 22 months in total, the first and last figures contain only two months.

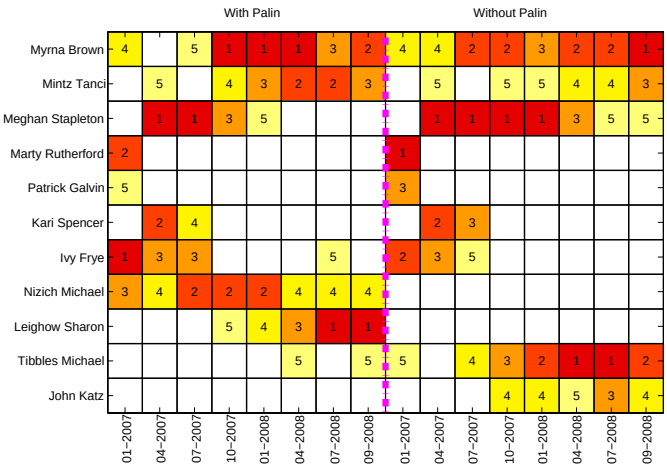


Figure 4: Evolution of Key Individuals' Social Statuses With & Without Sarah Palin

We observe different temporal patterns of the key individuals from Figure 4:

- **Persistent** Persistent key individuals steadily maintain their high social status for a long time. Also, the removal of Palin causes little impact to them, like Myrna Brown and Mintz Tanci.
- **Transient** Transient key individuals have high social status for a very short time period, like Marty Rutherford and Kari Spencer.
- **Active-With-Palin** These individuals maintain high social status in the networks with Palin, but they are not important after the removal of Palin, like Leighow Sharon and Nizich Michael.
- **Active-Without-Palin** They do not have high social statuses in the networks with Palin, but become more important after the removal of Palin, like Tibbles Michael and Jonh Katz.

Persistent individuals are more important than others as their social statuses are consistently high along the timeline. As we can see, many of them are also top key individuals in Section 3.1. Transient individuals have high social status during a specific time period. They are involved in some social events and become key individuals because of their distinct social role. Active-with-Palin and active-without-Palin individuals have their distinct social status in the social network and cannot be easily detected. The Active-with-Palin individuals have many connections with Palin and they can be recognized as the people who communicate with Palin directly. For example, Leighow Sharon is a spokeswoman for Palin. She is very important in the network with Palin, but significantly less important after Palin is removed. She has many communications directly with Palin, but does not help Palin manage specific tasks. The active-without-Palin individuals appear important after the removal of Palin, it further demonstrates that they can be recognized as "alternatives" of Palin. Without temporal features and the "knockout test", it is impractical to identify the transient

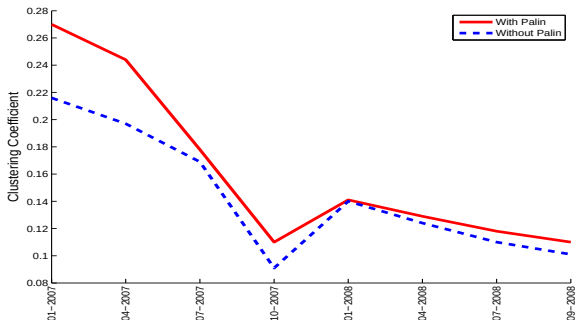


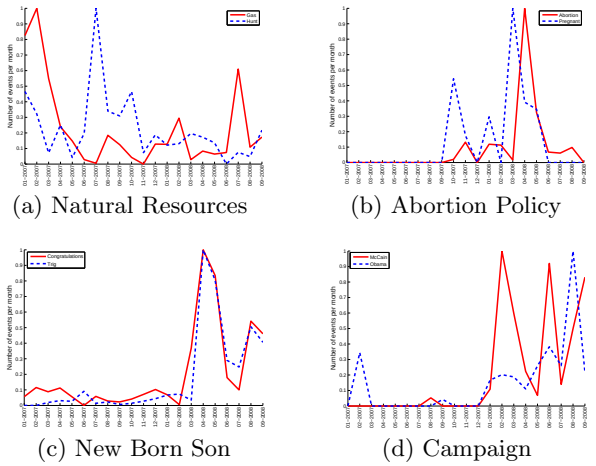
Figure 5: Clustering Coefficient Evolution With & Without Sarah Palin

key individuals and Active-Without-Palin individuals, who also have high social statuses in the network.

3.3.2 Network Structure Evolution

Figure 5 depicts the clustering coefficient result of the eight sub-networks at different points in time. In the figure, a red continuous curve represents the sub-networks with Palin and a blue dotted curve indicates sub-networks after knocking out Palin. The difference between the two figures is large at 01-2007, indicating the network structure has significant change between the two sub-networks at that time. The difference between two curves becomes smaller along the timeline, which indicates the structure between two sub-networks becomes more similar. From this figure, it is clear that the removal of Palin brings in different impact to the networks at different time points.

To further explore the rationale behind the different impact brought by the knocking out of Palin at different time, we depict the networks with and without Palin at the time points in Figure 6. Figure 6(a1) to 6(a8) represent the networks with Palin at eight time points, and Figure 6(b1) to 6(b8) depict the networks after the removal of Palin at the same time points. There is one clear center (Palin) in Figure 6(a1). Key individuals consist of a core in the networks from Figure 6(a2) to 6(a3). From Figure 6(a4) to 6(a7), it appears that many weakly mutually connected people connect to only one key individual and they consist of some small groups around the network. For the ease of presentation, we call the center individuals of the small groups *distinct connected individuals* in this paper. These distinct connected individuals have high betweenness centrality and many marginal individuals connect to this network via them. Through the networks with and without Palin in Figure 6(a1) and 6(b1), the network is destroyed significantly by removing Palin. It is because many people are only connected to Palin, but not any other ones in the network. Although shape of the networks in Figure 6(b2) to 6(b3) are similar as that in Figure 6(a2) to 6(a3), the communications within the core significantly decrease and many nodes become isolated. It is because some nodes are only connected to Palin and the network structure is not reliable. From Figure 6(b4), although the internal communications decrease, the network structure remains robust. This is because there are more distinct connected individuals, who play an important role to maintain the network



(a) Natural Resources (b) Abortion Policy

(c) New Born Son (d) Campaign

Figure 7: Event Evolution from 01-2007 to 09-2008

structure even after the removal of Palin.

The rationale behind this is that many new members joined the network during this time period. As we discussed in Section 2.1, at that time, Palin was involved in many events, like her daughter's pregnancy and marriage, she was announced as John McCain's running mate etc. Thus her social activities and communication increased during the period.

3.3.3 Event Evolution

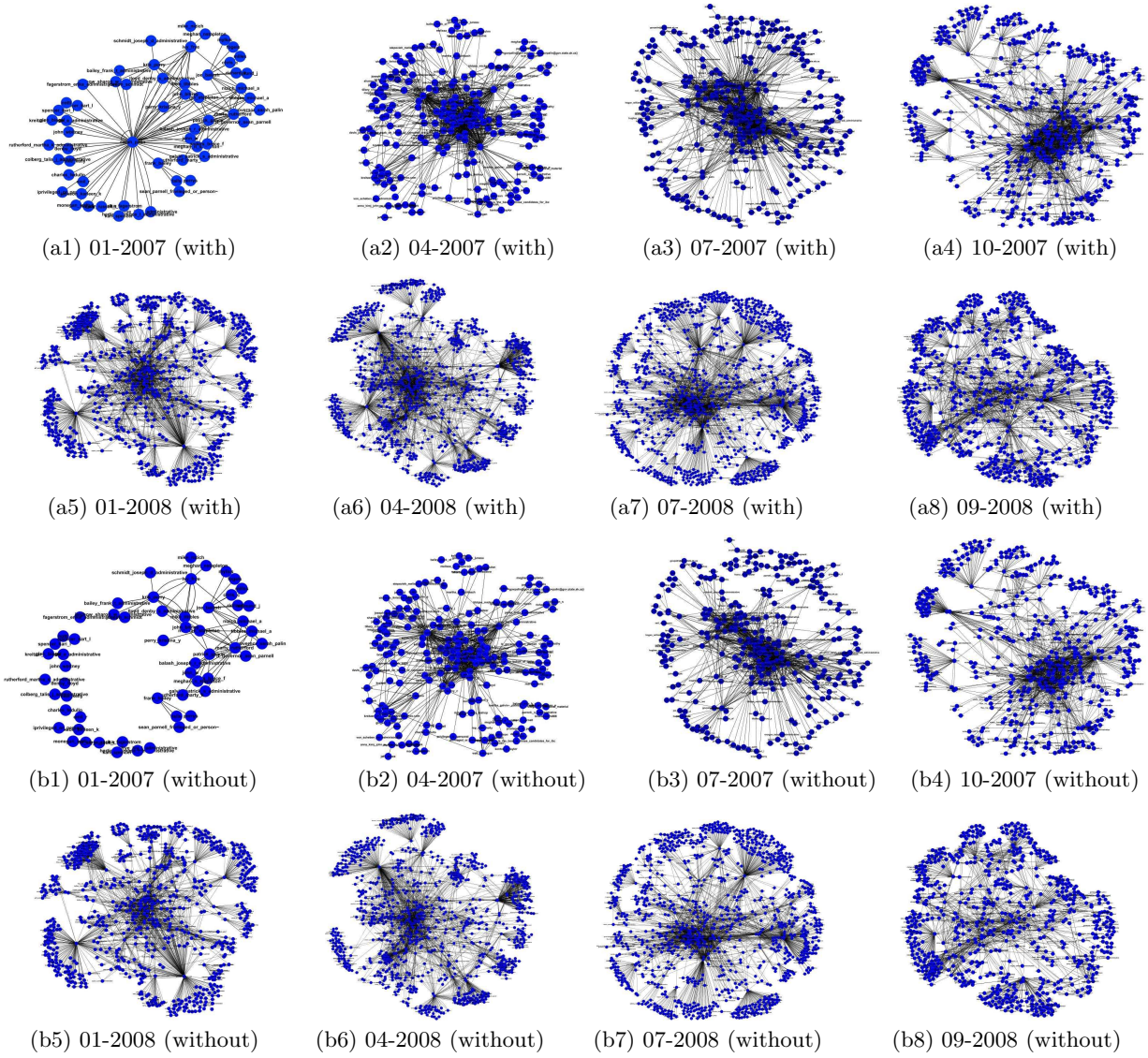
As we discussed in Section 3.3.1 and Section 3.3.2, there are some transient key individuals appearing and the groups members change significantly during some specific time period. To further explore particular impact during these time periods and verify our finding, we investigate evolution of some important events, the normalized results are shown in Figure 7. Four kinds of events are carefully selected, including "Natural Resources", "Abortion Policy", "New Born Son" and "Campaign", from the "Key Events Timeline" published by The Guardian[9].

Figure 7(a) depicts the results for "Gas" and "Hunt". The curve is very active in 2007 and peaks at 02-2007 and 07-2007. It is because Sarah Palin supported the wolf hunting policy and was selected as the head of the Alaska Oil and Gas Conservation Commission. Also, we discussed transient key individuals in Section 3.3.1 and both sample individuals are all from the Department of Natural Resources.

In Figure 7(b) to 7(d), we can see the events are very active during 02-2008 to 07-2008. This is because events related to Sarah Palin occurred during that period. The increase in activity is likely attributed to the events relating to Palin including the pregnancy of her daughter and becoming McCain's running mate.

These public events also caused Palin's network to significantly change at that time, which is evidenced by Figure 8. In the Figure, a red continuous curve represents distinct individuals each month and a blue dotted curve indicates the stable individuals who also appeared in the previous month. The dotted curve is more stable than the continuous one, which means that the individuals changed significantly be-

[9]http://www.guardian.co.uk/world/interactive/2011/jun/11/sarah-palin-emails-interactive-timeline



(a1) 01-2007 (with) (a2) 04-2007 (with) (a3) 07-2007 (with) (a4) 10-2007 (with)

(a5) 01-2008 (with) (a6) 04-2008 (with) (a7) 07-2008 (with) (a8) 09-2008 (with)

(b1) 01-2007 (without) (b2) 04-2007 (without) (b3) 07-2007 (without) (b4) 10-2007 (without)

(b5) 01-2008 (without) (b6) 04-2008 (without) (b7) 07-2008 (without) (b8) 09-2008 (without)

Figure 6: Network Evolution With and Without Sarah Palin

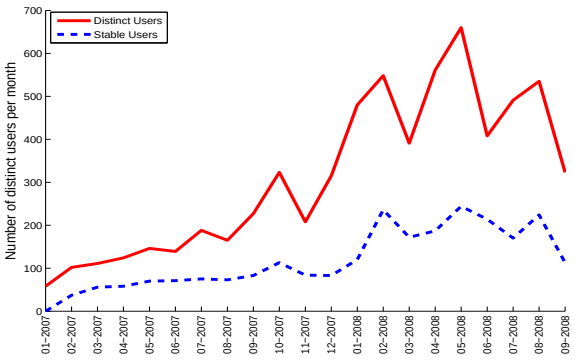


Figure 8: Group Members Evolution

tween two months, but some appear continuously active. The continuous line reaches the peak in 05/2008. It appears active and individuals in the network change significantly during 02/2008 to 06/2008. These changes impact Palin's network composition. More key individuals help Palin receive and spread information in the network, but not many people connect to Palin directly. This change makes network centralization decrease and the social structure more reliable.

4. RELATED WORK

Recently, online social network analysis has gained huge popularity and attracted researchers from disciplines. Email communication is an early form of social media. Although email service has been available for many years, email network analysis is also in early stages due to the lack of significant data. To the best of our knowledge, this is the first look at Sarah Palin's email corpus. We provide a compre-

hensive solution to analyze a personal email network. There are, however, several lines of related work.

Social Status and Role Analysis

Social status can be considered the relative rank that an individual holds in a social hierarchy based upon honor or prestige. Some methods have been proposed to find key (a.k.a. important or influential) individuals with high social status in various social networks. Newman [19] proposed to capture the strengths of collaborative ties to find the best connected scientist in scientific coauthorship networks; Different features of influential individuals in blogosphere [1] and microblogosphere [29] are investigated. Our paper does not focus on the impact of different pre-defined features, but seeks to understand the people with high social status from different perspectives.

To better understand the behavior and motivations of different types of people, identifying social roles in a community has been studied for years. Topics are extracted from documents and studied to find corresponding social roles [16, 18]; Thom-Santelli et al [26] employed the concept of social roles to analyze audience-oriented tagging, including roles of evangelist, publisher, and team-leader.

Email Analysis

Investigations on email corpus have been conducted by researchers for years, which can generally be grouped into two categories, personal and organizational email network analysis [23]. Before Sarah Palin's email corpus was made available, research on personal email network analysis more used synthetic or toy datasets to do verification [23]. In contrast, Enron's email corpus [13] was made public during the legal investigation concerning Enron corporation, allowing for organizational email network analysis to make great progress. Although many publications focused on Natural Language Processing and Text Mining [3], the explorations on the Enron email corpus are from different aspects, including socio-cognitive analysis [21], structural exploration [6], and community evolution analysis [24].

There are significant differences between organizational and personal email networks. Organizational email communications are strictly monitored by the company, the social hierarchy is clear and the social roles are fixed [22]. In personal email networks, the communications are more casual and the actors might have multiple social roles. In our case, Sarah Palin communicates with people from the governor's office, family members and her friends.

Online Social Network Analysis

The rising popularity of online social media services has spurred research of online social network analysis into two main directions. One way is to study the characteristics or verify social theory by exploring some social media data sets. Researchers studied the topological characteristics [14] and media communication [30] on Twitter. Yang and Leskovec [32] aim to study temporal patterns associated with online content and how the content's popularity grows and fades over time. The other is to improve variant applications in social media via social network analysis. Some related applications include identifying the influential bloggers [1] and twitterer [29], link prediction [2], election prediction [27] etc.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we address the challenges in examining Palin's email documents. We first employ conventional text analytical tools to reconstruct Palin's email network. We conduct a detailed study on social statuses and social roles of Sarah Palin's emails from three perspectives (individual, group, and temporal). With our text analytic tool, we draw interesting findings: there are many other important people playing different roles based on our analysis; as time passes by, the removal of Palin from the email tends not to cause significant negative effects on the network, although she has the highest social status; some key individuals with high social statuses are thus identified to help understand their changing relationships with Palin's various activities. In the temporal analysis, we find who are the key individuals in different phases and validate these different temporal patterns with the event evolution. Our proposed analytic tool is designed to help the public to effectively and efficiently examine the email corpus, navigate complicated relationships, and investigate overlapping groups and their evolution, otherwise either impractical or difficult.

There are a number of interesting extensions of this work. Our proposed analytic tool can be further developed to study similar egocentric social networks. The Sarah Palin's email corpus can be made available for the community to collaboratively explore various text analytics applications, like event detection, topic evolution, and group analysis.

6. REFERENCES

[1] N. Agarwal, H. Liu, L. Tang, and P. Yu. Identifying the influential bloggers in a community. In *Proceedings of WSDM*. ACM, 2008.

[2] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of WSDM*, pages 635–644. ACM, 2011.

[3] R. Bekkerman, A. McCallum, G. Huang, et al. Automatic categorization of email into folders: Benchmark experiments on enron and sri corpora. *Center for Intelligent Information Retrieval, Technical Report IR*, 418:1, 2004.

[4] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. In *Proceedings of WWW*, volume 7, pages 757–786, 2007.

[5] H.-H. Chen, M.-S. Lin, and Y.-C. Wei. Novel association measures using web search with double checking. In *Proceedings of the 21st COLING and the 44th ACL*, pages 1009–1016, 2006.

[6] J. Diesner and K. Carley. Exploration of communication networks from the enron email corpus. In *Proceedings of Workshop on Link Analysis, SDM*, pages 21–23. 2005.

[7] T. Egener, J. Granado, and M. Guitton. High frequency of phenotypic deviations in physcomitrella patens plants transformed with a gene-disruption library. *BMC Plant Biology*, 2(1):6, 2002.

[8] L. Garton, C. Haythornthwaite, and B. Wellman. Studying online social networks. *Journal of Computer-Mediated Communication*, 3(1):0–0, 1997.

[9] A. Giddens, M. Duneier, and R. Appelbaum. *Introduction to sociology*. Norton, 1991.

[10] A. Hollingshead. *Four factor index of social status*. Yale Univ., Dep. of Sociology, 1975.

[11] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing

the spread of influence through a social network. In *Proceedings of SIGKDD*, pages 137–146. ACM, 2003.

[12] J. Kleinberg and S. Lawrence. The structure of the web. *Science*, 294(5548):1849, 2001.

[13] B. Klimt and Y. Yang. The enron corpus: A new dataset for email classification research. *Machine Learning: ECML 2004*, pages 217–226, 2004.

[14] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[15] J. Leskovec, K. Lang, and M. Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of WWW*, pages 631–640. ACM, 2010.

[16] A. Leuski. Email is a stage: discovering people roles from email archives. In *Proceedings of SIGIR*, pages 502–503. ACM, 2004.

[17] G. Manco, E. Masciari, M. Ruffolo, and A. Tagarelli. Towards an adaptive mail classifier. In *Proc. of Italian Association for Artificial Intelligence Workshop*. Citeseer, 2002.

[18] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pages 786–791. 2005.

[19] M. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. *Complex networks*, pages 337–370, 2004.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[21] N. Pathak, S. Mane, and J. Srivastava. Who thinks who knows who? socio-cognitive analysis of email networks. In *Proceedings of ICDM*, pages 466–477. IEEE, 2006.

[22] R. Rowe, G. Creamer, S. Hershkop, and S. Stolfo. Automated social hierarchy detection through email network analysis. In *Proceedings of WebKDD and SNA-KDD workshop*, pages 109–117. 2007.

[23] M. Smith, B. Shneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave. Analyzing (social media) networks with nodexl. In *Proceedings of the fourth international conference on Communities and technologies*, pages 255–264. ACM, 2009.

[24] L. Tang, H. Liu, J. Zhang, and Z. Nazeri. Community evolution in dynamic multi-mode networks. In *Proceeding of SIGKDD*, pages 677–685. ACM, 2008.

[25] L. Tang, X. Wang, H. Liu, and L. Wang. A multi-resolution approach to learning with overlapping communities. In *Proceedings of the First Workshop on Social Media Analytics*, pages 14–22. ACM, 2010.

[26] J. Thom-Santelli, M. Muller, and D. Millen. Social tagging roles: publishers, evangelists, leaders. In *Proceeding of SIGCHI*, pages 1041–1044. ACM, 2008.

[27] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of ICWSM*, pages 178–185, 2010.

[28] S. Wasserman. *Social network analysis: Methods and applications*. Cambridge university press, 1994.

[29] J. Weng, E. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of WSDM*, pages 261–270. ACM, 2010.

[30] S. Wu, J. Hofman, W. Mason, and D. Watts. Who says what to whom on twitter. In *Proceedings of WWW*, pages 705–714. ACM, 2011.

[31] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *Proceedings of WWW*, pages 981–990. ACM, 2010.

[32] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of WSDM*, pages 177–186. ACM, 2011.

APPENDIX

In Table 7, we summarize all the important individuals mentioned in this paper and their corresponding positions. The positions of the people are manually verified from news articles and online resources, and grouped into six categories, including personal assistants, campaign aides, family members, people from outside company, staff members from different departments of the state of Alaska, and "not clear" if no position found via search. Clearly, most of the individuals discovered with our analytic tool play important roles in Palin's social network. It further demonstrates the effectiveness of our analytic tool in examining Palin's email corpus. Among the key individuals, it is noted that some of them have no obvious public positions. It is interesting to probe further who they really are, what made them so important. Our analytic tool can be used by the general public in assisting them to examine many email documents and various relationships.

Table 7: Key individuals and Positions

Key individuals	Positions
Myrna Brown	Assistant
Janice Mason	Scheduler
Kari Spencer	Scheduler
Ivy Frye	Campaign aide
John Katz	Head of Washington office
Leighow Sharon	Spokesperson
Meghan Stapleton	Spokesperson
Bristol Palin	Eldest daughter
Todd Palin	Husband
Track Palin	Eldest son
Kevin Haper	Consultant at Black & Veatch's
Anders Bruce	Department of Natural Resources
Josheph Balash	Energy Advisor
Patrick Galvin	Department of Revenue
Beth Leschper	State of Alaska
Tibbles Michael	Chief of staff
Kristina Perry	Director of Anchorage Office
Martha Rutherford	Department of Natural Resources
Kim Anna	Not Clear
Ausman Earle	Not Clear
Nizich Michael	Not Clear
Mintz Tanci	Not Clear