24,000 pages of emails were then made available by CNN[2]. To process the corpus, we employ Optical Character Recognition software provided by Adobe Acrobat to convert all images to text documents. To retain important information, each email is indexed as: $E = <$*"Body", "Subject", "Sent", "From", "To", "CC", "Size", "numRecipients"*$>$. Finally, all the emails are processed in an XML corpus.

2.2 Name Entity Resolution

The quality of generated name phrases is low, and a person intentionally 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 method is based on the hypothesis that when two phrases have more words that overlap, they are considered to be related more.

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. 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 – Jaccard coefficient, is used as below:

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

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

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

Thus, the similar name phrases are mapped into distinct name entities.

2.3 Email Network Reconstruction

In the language of email communication network, entities correspond to people sent or received email, and edges correspond to relations sent-by. We reconstruct email network based on <"From", "To", "CC"> fields of our built XML corpus. A list of user names extracted from the emails is processed by our proposed name entity resolution methods. The output identified entities are employed as vertices of the network. Edges are added between pairs of entities (sender and receiver). The graph edges are directed from the sender to the receiver of the email. We do not add edges from a node to itself. The link weight between two nodes is represented as the number of emails sent and received between two people.

3. STATUS & ROLE ANALYSIS

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. The network centralization of 0.220 represents this network is not strictly centered by one node as a star structure [5].

Table 1: Networks With & Without Sarah Palin

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

We define the people with high social status as *key individuals* in the social network. 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 to test the overall performance variance brought by one process or one component when it is made inoperative in the framework [1]. We conduct 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.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we employ text analytical tools to reconstruct Palin's email network, and provide a preliminary study of Palin's social status and social role in the network [4]. Also, the refined Sarah Palin's email corpus could be useful to collaboratively explore various text analytics applications, like event detection, topic evolution, and group analysis.

5. ACKNOWLEDGMENTS

This work is, in part, supported by ONR.

6. REFERENCES

[1] 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*, 2002.

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

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

[4] X. Hu and H. Liu. Social status and role analysis of palin's email network. Technical report, Arizona State University, 2011.

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

[2]http://www.cnn.com/specials/2011/palin.emails/index.html