Analysing Information Flows and Key Mediators through Temporal Centrality Metrics

John Tang
University of Cambridge

Mirco Musolesi[*]
University of St. Andrews

Cecilia Mascolo
University of Cambridge

Vito Latora
University of Catania

Vincenzo Nicosia
University of Catania

ABSTRACT

The study of influential members of human networks is an important research question in social network analysis. However, the current state-of-the-art is based on static or *aggregated* representation of the network topology. We argue that dynamically evolving network topologies are inherent in many systems, including real online social and technological networks: fortunately the nature of these systems is such that they allow the gathering of large quantities of fine-grained temporal data on interactions amongst the network members.

In this paper we propose novel *temporal* centrality metrics which take into account such dynamic interactions over time. Using a real corporate email dataset we evaluate the important individuals selected by means of static and temporal analysis taking two perspectives: firstly, from a semantic level, we investigate their corporate role in the organisation; and secondly, from a dynamic process point of view, we measure information dissemination and the role of information mediators. We find that temporal analysis provides a better understanding of dynamic processes and a more accurate identification of important people compared to traditional static methods.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Network Topology; C.2.0 [**General**]: Data communications

General Terms

Measurement, Algorithms, Theory

---
[*]This work was carried out in part when Mirco Musolesi was at the Computer Laboratory, University of Cambridge.

Keywords

Temporal Graphs, Temporal Metrics, Temporal Centrality, Key Mediators, Social Networks, Complex Networks, Information Diffusion

1. INTRODUCTION

Identifying key nodes has become an essential part of analysing and understanding networked systems with application to a wide range of fields including finding the best person to target in a viral marketing campaign [15, 22], locating key neurons in cortical networks [5], protecting important species in ecological systems [14] and finding bottlenecks in traffic networks [13]. The position of a node with respect to other nodes can be classified and exploited: one could argue that people with the most friends are popular and hence important; a node with high geodesic locality to other nodes could spread information quickly to high numbers of nodes; and a person who lies between the most paths of communication could act as a mediator among groups of people. These concepts are more commonly known as degree, closeness and betweenness centrality [22, 4].

Existing centrality metrics for the study of real online social networks (OSN) are based on a static network model where edges that appear (and disappear over time) are aggregated into a single static graph [1, 9, 10]. In particular, if we consider a graph of interactions over time where each edge corresponds to an interaction (such as posting of a message) between two users represented by the nodes, we observe a continuous change of the graphs corresponding to the dynamic user activity. Examples are instant messenger systems and email where also the interactions often happen in bursts [17].

With this in mind, the key contribution of this paper is the introduction of *temporal centrality metrics* for the identification of key nodes in OSNs based on temporal shortest paths. Tang et al. [20] proposed a *temporal network* model to study real dynamic dataset. The idea is that the behaviour of dynamic networks can be more accurately captured by a sequence of *snapshots* of the network topology as it changes over time (Figure 1). Using this model, a temporal version of shortest path was also defined. Since static closeness and betweenness centrality metrics are derived from static shortest paths, we extended these metrics to temporal graphs by introducing the definitions of *temporal closeness* and *temporal betweenness*. Naturally, both these temporal extensions are
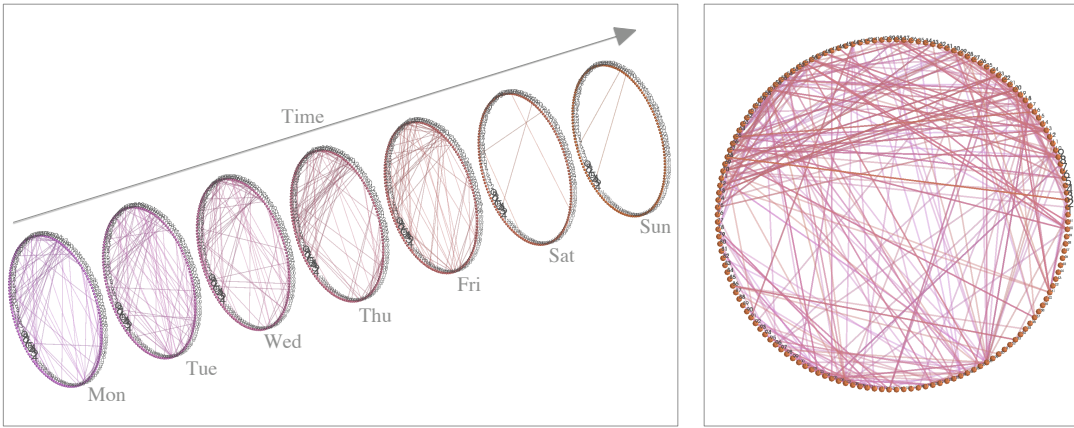
Figure 1: Temporal Graph showing a typical week of activity during Nov 2001 using 24-hour windows (left) and aggregated static graph (right). Nodes represent employees; a link between two employees exists if an email was sent by one of them to the other in that 24-hour window.

associated to the identification of central nodes in the network with application to *dynamic processes* over an OSN. In particular, temporal closeness quantifies how fast a user can disseminate a piece of information. Therefore, applications of this metric include viral marketing and the study of rumour spreading. On the other hand, temporal betweenness distinguishes individuals who act as key *mediators* between the most communication paths over time.

Hence, to evaluate our proposed temporal centrality metrics, we apply them to separate dynamic processes. First, an information dissemination process captures the notion of speed and reachability of a piece of information spreading through a network starting from a subset of nodes. Thus, we can compare the important nodes selected by static and temporal closeness centrality. Second, since we observe that if certain individuals lie between the most communication paths, their removal from the network will impact the overall communication efficiency.

We provide a preliminary evaluation using the publicly available Enron email dataset [19], describing corporate communication over time between 151 known employees during the height of the company's accounting scandal.

The main contributions of this paper are as follows:

- Based on temporal path lengths proposed by Tang et al. [20] we define novel temporal centrality metrics for the study of key nodes in OSNs (Section 2).

- We evaluate these temporal metrics using two dynamic processes, applied on the Enron email dataset. The temporal analysis consistently finds that employees who worked as energy traders fulfil both these roles. This gives us an interesting insight not only into the Enron scandal but also into the differences between static and temporal analysis (Section 3.3). We also find that compared to existing static analysis, temporal metrics not only uncover important nodes that are better for information spreading but also individuals who play a vital role in mediating between the most communication channels (Section 3.4).

- We provide initial insights into temporal dynamics which make temporal metrics more suitable for studying time-varying OSNs compared to static analysis (Section 3.5).

2. TEMPORAL CENTRALITY METRICS

As discussed, the representation of a time-varying network by means of the associated static graph can convey misleading information about the network itself. For instance, a static aggregated network usually has far more links than the network has at each time instant (or if the aggregation is performed over short time windows). Since the time ordering of edges is not considered, the number of paths between pairs of nodes is overestimated. This also leads to an underestimation of the length of shortest paths [20].

Following from this, since traditional static closeness and betweenness centrality are based upon static shortest paths, the estimation provided by these metrics is largely inaccurate. In order to overcome these problems, in this section, we introduce the definitions of *temporal closeness* and *temporal betweenness* centrality by employing temporal shortest paths, which do take into account time information. First we will present the basic definitions of the temporal graph model and temporal paths.

2.1 Preliminaries

A *temporal graph* can be thought of as an ordered sequence of graphs. A state of the network topology is calculated by aggregating all the edges that appear inside a certain time window. An example is given in Figure 1: the temporal graph shown in the left panel is a sequence of seven graphs, each of them representing the contacts among nodes in a time window of 24-hours. The corresponding aggregated static graph (which reports all the links among nodes, without any information about time) is shown in the right panel.

More formally, given a real network trace starting at t_{min} and ending at t_{max}, the temporal graph $\mathcal{G}_t^w(t_{min}, t_{max})$ is defined as the ordered sequence of graphs $G_{t_{min}}$, $G_{t_{min}+w}$, ..., $G_{t_{max}}$, where w is the size of each time window, expressed in some time units (e.g., seconds or hours). The number of graphs in the sequence is denoted as $W = ((t_{max} - t_{min})/w) = |\mathcal{G}_t^w(t_{min}, t_{max})|$. The contact function R_{ij}^s between nodes i, j at time s is equal to 1 if and only if there exists a link between i and j in $G_t, t \leq s \leq t + w$, otherwise R_{ij}^s is equal to zero. All the graphs in the temporal graph have the same set of nodes V, while each of them has, in general, a different set of edges E_t, where an edge between

$i, j \in V$ belongs to E_t if and only if $R_{ij}^s = 1$. From this model a *temporal path* between two nodes i and j can be defined over $\mathcal{G}_t^w(t_{min}, t_{max})$ as a sequence of k hops via a distinct node $n_k^{t_k}$ at time t_k:

$$p_{ij}^h = (n_1^{t_1}, \ldots, n_k^{t_k}) \qquad (1)$$

starting from node $i = n_1$ and finishing at node $j = n_k$, where $t_{k-1} \leq t_k$, node n_k is passed a message at time window $t_{min} \leq t_k \leq t_{max}$ and h is the maximum number of exchanges between nodes within the same window t. We call Q_{ij}^h the set of all temporal paths between nodes i and j. If a temporal path from i to j does not exist i.e. $Q_{ij}^h = \emptyset$, we say that (i, j) is a *temporally disconnected node pair*, and we set the distance $d_{ij} = \infty$.

Using the function $D(p_{ij}) = t_k - t_{min}$ which returns the delivery time for the given path relative to t_0, the *shortest temporal path length* is defined as:

$$d_{ij}^h = \forall q_{ij} \in Q_{ij}^h, min(D(q_{ij})) \qquad (2)$$

From this we define the set S_{ij} of *shortest temporal paths* between i, j as:

$$S_{ij}^h = \{p_{ij}^h \in Q_{ij}^h \mid D(p_{ij}^h) = d_{ij}^h\} \qquad (3)$$

Notice also that, in general, d_{ij}^h is different from the length of the *physical path* from i to j, expressed as the number of links to be traversed to reach j starting from i, since in a temporal path the message can be passed to up to h subsequent nodes within the same time window. The *horizon* of the temporal graph models the speed of message passing between nodes compared to the length of the time window.

2.2 Temporal Betweenness Centrality

Static betweenness centrality of a node i is defined as the fraction of shortest paths between all pairs of nodes which pass through i [22]. In principle, the temporal betweenness centrality of node i could be defined as the fraction of *temporal* shortest paths that pass through i. However, for the study of OSNs, identifying nodes which mediate communication between the most groups of nodes is of paramount importance. Such nodes represent members of society who both receive and are the source of gossip; individuals who negotiate between the different groups of parties; and people in organisations who fall into middle management and balance reporting to senior management and also command a large workforce. If such nodes provide an important mediatory role in a network then it stands that the complementary view would also hold; how does the removal of such nodes disrupt the overall communication efficiency of the network? It is therefore important to take into account not only the number of shortest paths which pass through a node, but also the length of time for which a node along the shortest path *retains* a message before forwarding it to the next node.

For example, let us consider the simple case of nodes i and j being connected by just one shortest path $p_{ij} = (i^{t_0}, k^{t_1}, j^{t_2})$, so that a message from i to j has to pass first through k at time t_1 before being delivered to j at time t_2. Since the path through k is the only way for i to send a message to j, then we would say that k plays an important mediatory role and is "central" for communication between i and j. Nevertheless, the vulnerability of node k heavily depends on the intervals $[t_0, t_1]$ and $[t_1, t_2]$: the longer the

time that a message forwarded from i has to wait on k before being passed to j, the higher the chance of disruption removing the message which was destined to j.

From this the *temporal betweenness centrality* of node i at time t is defined as:

$$C_i^B(t) = \frac{1}{(N-1)(N-2)} \sum_{\substack{j \in V \\ j \neq i}} \sum_{\substack{k \in V \\ k \neq i \\ k \neq j}} \frac{U(i, t, j, k)}{|S_{jk}^h|} \qquad (4)$$

defined when $S_{jk}^h \neq \emptyset$, where the function U returns the number of shortest temporal paths from j to k in which node i has either received a message at time t or is holding a message from a past time window until the next node is met at some time $t' > t$. In the case when $S_{jk}^h = \emptyset$, i.e., node i is totally isolated, we set its betweenness to zero. Finally, the temporal betweenness for node i over the entire temporal graph $\mathcal{G}(t_{min}, t_{max})$ is:

$$C_i^B = \frac{1}{W} \sum_{t=1}^{W} C_i^B((t \times w) + t_{min}) \qquad (5)$$

2.3 Temporal Closeness Centrality

Two nodes of a static graph are said to be *close* to each other if their geodesic distance is small. In a static graph an estimation of the global *closeness* of a node i is obtained as the average static shortest path length to all other nodes in the graph. Similarly, we can extend the definition of closeness to temporal graphs using the temporal shortest path length between nodes, which is a measure of how early a source node can deliver a message to all other nodes.

Given the shortest temporal distance $d_{ij}(t_{min}, t_{max})$, the temporal closeness centrality can be expressed as:

$$C_i^h = \frac{1}{W(N-1)} \sum_{j \neq i \in V} d_{i,j}^h \qquad (6)$$

so that nodes that have, *on average*, shorter temporal distances to the other nodes are considered more *central*.

3. EVALUATION

3.1 Enron Email Dataset

The Enron Energy Corporation started as a traditional gas and electrical utility supplier; however, in the late 1990s their main money making business came from trading energy on the global stock markets [11]. In December 2001, the Enron Energy Corporation filed for bankruptcy after it was uncovered that fraudulent accounting tricks were used to hide billions of dollars in debt [12]. This led to the eventual conviction of several current and former Enron executives [21, 7]. The investigation also brought to light the reliance of the company on traders to bring in profits using aggressive tactics culminating in intentional blackouts in California in Summer 2001. With both control over electricity plants and the ability to sell electricity over the energy markets, Enron traders artificially raised the price of electricity by shutting down power plants serving the State of California and profiting by selling electricity back at a premium [6].

During the investigation into the Enron accounting scandal, telephone calls, documents and emails were subpoenaed by the U.S. government and as such the email records of 151
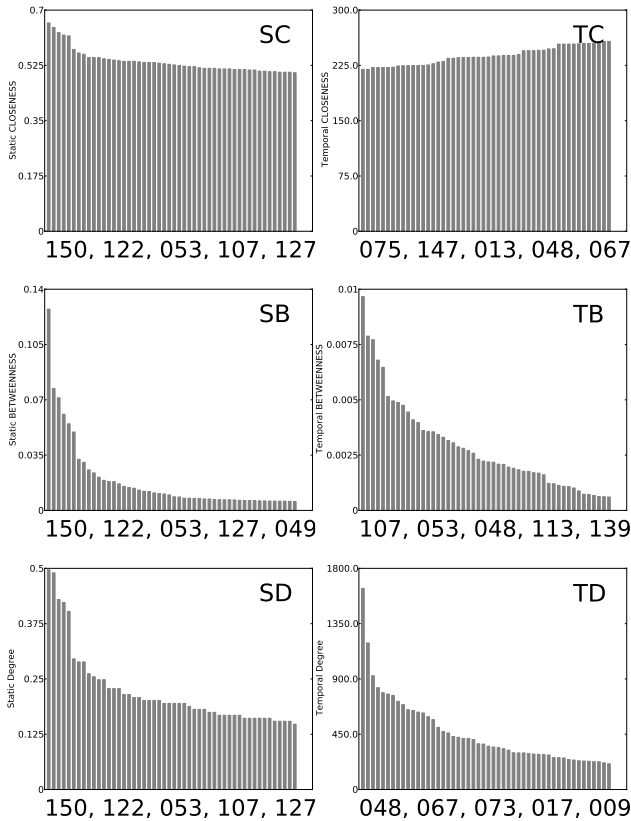
Figure 2: Ranked distribution of top 50 statically (S) and temporally (T) central nodes. From top row: Closeness (C), Betweenness (B), and Degree (D). Top 5 node ID's listed under each plot.

ID	Name	Role	Notes
9	Stephanie Panus	(Unknown)	
13	Marie Heard	Legal	Senior Legal Specialist
17	Mike Grigsby	Manager	
48	Tana Jones	Executive	
53	John Lavorato	Trader	
54	Greg Whalley	President	Former Head of Trading
67	Sara Shackleton	Vice President	Enron Wholesale Services
73	Jeff Dasovich	Trader	
75	Gerald Nemec	Director of Trading	
107	Louise Kitchen	Trader	Head of Online Trading
122	Sally Beck	Managing Director	
127	Kenneth Lay	Chairman & CEO	
139	Mary Hain	Director	
147	Carol Clair	Trader	
150	Liz Taylor	Secretary	Assistant to Greg Whalley

Table 1: Roles of top centrality nodes.

	SB	SC	SD	TB	TC	TD
SB	1.00	0.57	0.69	0.41	0.24	0.43
SC	-	1.00	0.70	0.36	0.22	0.31
SD	-	-	1.00	0.39	0.28	0.48
TB	-	-	-	1.00	0.43	0.34
TC	-	-	-	-	1.00	0.40
TD	-	-	-	-	-	1.00

Table 2: Kendall-tau correlation coefficients between centralities.

user mailboxes were part of the public record consisting of approximately 250,000 emails sent and received during the period between May 1999 to June 2002 (1137 days), leading up to the bankruptcy filing. None of the emails were anonymised and so they provide unique semantic information of the owner of each mailbox.

3.2 Temporal Graph Construction

In our analysis, we use the dataset prepared by Shetty & Adibi [19]. Since we do not have a complete picture of the interactions of users outside of the subpoenaed mailboxes we concentrate on email exchanges between the core 151 users only. Taking this email dataset, we process the complete temporal graph from 1999 to 2002 with undirected links, using windows of size w=24 hours and horizon h=1. If an email was exchanged between two individuals in a temporal window, a link between the two nodes representing those individuals will be added to the graph representing the temporal snapshot for that time.

3.3 Semantic Value of Temporal Centrality

Figure 2 plots the static and temporal centrality rankings of employees calculated using closeness and betweenness. Examining the static centralities (left column) we note that there is little difference between the top five employees using static closeness or betweenness. Also plotting the static

degree centrality of each node[1], we notice similar rankings suggesting that static analysis only favours employees who interacted with the most number of other people. Temporal closeness and temporal betweenness yield different rankings amongst the top five and the calculated Kendall-tau correlation coefficient[16] (Table 2) confirm that static-to-static metrics are strongly correlated ($\simeq 0.7$). Also note that there is low correlation (< 0.4) between temporal metrics and static degree demonstrating that temporal analysis is not dependent on the number of people an individual interacts with.

Cross referencing the top two employee identifiers with their position within the organisation (Table 1) we identify a secretary (150) and managing director (122) as central nodes for both static closeness and betweenness; however, both temporal closeness and betweenness consistently selected employees in trading roles (053, 075, 107, 147). A secretary and a managing director are certainly important for information dissemination and central to many communication channels, as detected by static measures. However, instead the top trading executives are exclusively favoured by temporal analysis. Moreover, cross-referencing with media reports [8], we find a correlation between the top two bonuses received and the two employees identified by temporal betweenness. To show that temporal analysis does not simply uncover nodes with the most interactions with other people, we also plot the temporal degree (TD) calculated as the total number of emails sent and received by each node i. Since there is a low correlation (< 0.4) with temporal closeness and betweenness this shows that temporal analysis is not dependent on the number of emails sent and received by each individual.

---

[1]The static degree centrality is defined as the number of edges connected to a node i, normalised by the total possible neighbour nodes $(n - 1)$ [22]
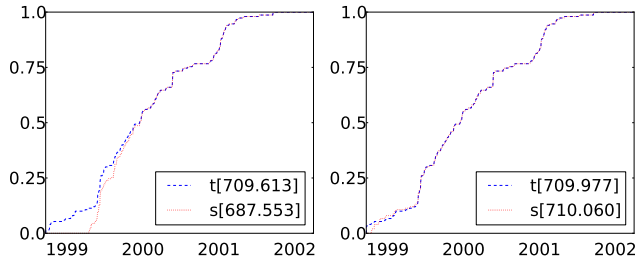
Figure 3: Dissemination Process: Dissemination ratio starting from top 2 (left) and top 10 (right) closeness source nodes. Area under curve reported in legend for temporal (t) and static (s) centrality.
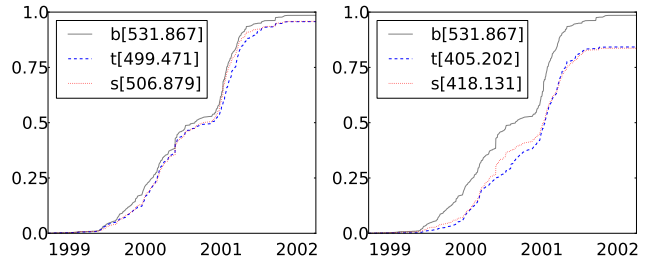

Figure 4: Mediation Process: Dissemination ratio after removing top 2 (left) and top 10 (right) betweenness nodes. Area under curve reported in legend for temporal (t), static (s) and baseline (b) where no nodes are removed.

3.4 Effectiveness of Central Nodes on Dynamic Processes

3.4.1 Trace-driven Simulation Setup

To evaluate the role and the centrality of the employees identified by temporal and static analysis, we consider two dynamic processes. First, we simulate a simple information *dissemination process* over the temporal graph constructed from the Enron traces. The process is simulated as follows. We select the top N nodes from the ranking based on temporal closeness centrality. We place an identical message m into their (infinite) buffers. We refer to any node that has received a copy of this message as *reached*. We then replay the contact trace through time and as reached nodes make contact with an unreached node u, the message is replicated into the buffer of node u. We assume that messages are transferred instantaneously and only the first neighbour in a time window can be reached. We then repeat this for static closeness centrality and plot the dissemination ratio across time for both.

Second, to model the role of individuals as part of an information *mediation process*, we borrow concepts from the more commonly known epidemic immunisation process where the dissemination ratio of a contagion spreading throughout a static network is measured before and after certain nodes are immunised against the contagion [2]. This is analogous to measuring the spread of information (the contagion) before and after important individuals are removed from the network (such as going on holiday or being discharged) since our conjecture is that removing mediators will impact the network communication efficiency greatly.

In the trace-driven simulation, instead of a single message spreading within the organisation, we seed all employees with a different message that needs to be delivered to all other employees. This models multiple channels of communication. In order to derive a baseline performance, we start by calculating the dissemination ratio when no nodes are removed. We then remove the top N individuals identified by temporal betweenness and rerun the information spreading process. Nodes which are removed cannot receive or pass on messages. We then repeat the same process for comparison using static betweenness centrality for the ranking.

3.4.2 Evaluating Information Dissemination & Mediation

We present plots using $N = \{2, 10\}$ for information dissemination (Figure 3) and information mediation (Figure 4).

As we can see the different pairs of traders identified by temporal analysis are better than the arbitrary nodes selected by static analysis for both disseminating information through the organisation and acting as mediators between communication channels. In the information dissemination case, although the final dissemination is the same across the long period of time, the two traders selected by temporal analysis disseminate information quicker. Only after increasing to 10 nodes the static analysis presents similar results. In the information mediation case, the final dissemination ratios for both temporal and static centrality nodes slightly decreases by removing the nodes but are comparable. However, removing the two traders gives an overall more prolonged drop in message dissemination. In the case of the removal of 10 nodes, the individuals identified by means of the temporal metrics slow the dissemination process further compared to static ones.

3.5 Insights into Temporal Dynamics

To gain some insight into the temporal dynamics of the individuals selected by temporal and static analysis, Figure 5 plots the number of emails sent and received over time, again by the top two centrality nodes. Also we recall from Section 3.4 that there is a strong correlation between static closeness and betweenness with degree. Such strong correlation between static closeness and betweenness with degree has been well documented in [18, 3].

By comparing the contact distribution between static analysis (top row) and temporal analysis (bottom row), we observe that the traders identified as important individuals by temporal analysis clearly interact earlier in time compared with the nodes identified by static analysis. This fits the intuition that earlier interactions are key to faster dissemination and hence temporal metrics are more accurate at identifying key individuals. This also confirms our arguments that static analysis ignores time information such as duration, frequency, time ordering and, at the simplest level, earlier interactions. Instead for both betweenness and closeness centrality, static analysis has prioritised individuals who interact with the highest number of different people.

4. CONCLUSIONS

In this paper we have presented novel temporal centrality metrics, namely temporal closeness and temporal betweenness to infer individuals playing a central role in time-
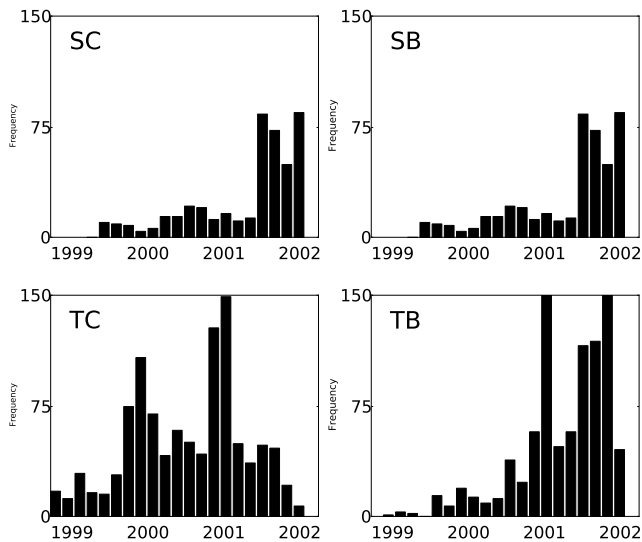
Figure 5: Distribution of total emails sent & received over time by top 2 centrality nodes. Bin size=50 days. From top-left: Static Closeness (SC), Temporal Closeness (TC), Static Betweenness (SB), Temporal Betweenness (TB).

varying networks from a communication perspective. More specifically, we have shown that by means of these metrics, we are able to identify nodes that are central in information dissemination processes with greater accuracy with respect to the corresponding static ones. We have evaluated and compared the static and temporal metrics by means of a large corporate email dataset. We have shown that these metrics are more effective for the identification of nodes that are able to speed up or block information diffusion processes in this class of networks.

These metrics are applicable to a number of system traces with timestamp information and, in general, to dynamic networks that change over time. We are in the process of further evaluating the potential of these modelling techniques and metrics in other contexts.

5. REFERENCES

[1] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *PNAS*, 101(11):3747–3752, Mar. 2004.

[2] A. Barrat, M. Barthélemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, Nov. 2008.

[3] M. Barthélemy. Betweenness centrality in large complex networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):163–168, Mar. 2004.

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, Feb. 2006.

[5] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.*, 10(3):186–198, Mar. 2009.

[6] CBS News. Enron traders caught on tape. http://www.cbsnews.com/stories/2004/06/01/ eveningnews/main620626.shtml, June 2004.

[7] CBS News. Former Enron Trader Pleads Guilty. http://www.cbsnews.com/stories/2004/06/16/ national/main623569.shtml, Aug. 2004.

[8] CNN. Enron paid hefty bonuses before bankruptcy. http://archives.cnn.com/2002/LAW/02/09/ enron.bonuses/index.html, Feb. 2002.

[9] E. Daly and M. Haahr. Social network analysis for information flow in disconnected Delay-Tolerant MANETs. *IEEE Transactions on Mobile Computing*, 8(5):606–621, 2009.

[10] H. Ebel, L. Mielsch, and S. Bornholdt. Scale-free Topology of E-mail Networks. *Phys. Rev. E*, 66(3):035103, 2002.

[11] P. Elkind and B. McLean. *The Smartest Guys in the Room: The Amazing Rise and Scandalous Fall of Enron*. Penguin, Sept. 2004.

[12] Federal Energy Regulatory Commission. Addressing the 2000-2001 Western Energy Crisis, Dec. 2008.

[13] P. Holmes. Congestion and centrality in traffic flow on complex networks. *Advances in Complex Systems*, 6(2):163–176, 2003.

[14] F. Jordan, T. A. Okey, B. Bauer, and S. Libralato. Identifying important species: Linking structure and function in ecological networks. *Ecological Modelling*, 216(1):75–80, Aug. 2008.

[15] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of ACM SIGKDD '03*, pages 137–146, 2003.

[16] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[17] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of SIGKDD '02*, pages 91–101, Edmonton, Alberta, Canada, 2002. ACM.

[18] M. J. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, 2005.

[19] J. Shetty and J. Adibi. Discovering important nodes through graph entropy the case of Enron email database. In *Proceedings of the 3rd International Workshop on Link Discovery*, pages 74–81, Chicago, Illinois, 2005. ACM.

[20] J. Tang, M. Musolesi, C. Mascolo, and V. Latora. Temporal distance metrics for social network analysis. In *Proceedings of WOSN '09*, 2009.

[21] Washington Post. Enron fraud trial ends in 5 convictions. http://www.washingtonpost.com/wp-dyn/articles/A23034-2004Nov3.html.

[22] S. Wasserman and K. Faust. *Social Networks Analysis*. Cambridge University Press, 1994.