From *SIAM News*, Volume 44, Number 2, March 2011

People Who Read this Article Also Read . . . : Part II

By Desmond J. Higham, Peter Grindrod, and Ernesto Estrada

Ideas from network science, having forced themselves the wrong way down Linton Freeman's one-way street [8] into the natural and engineering sciences, met a roundabout that speedily returned them to the realm of human social interactions. Many of the most important and exciting data sets of current interest to high-profile and high-spending users concern our behaviour: e-mailing, cell phoning, text messaging, satellite positioning, online socialising, online spending, movie renting. . . . Indeed, the most recent of the annual Network Science Conferences (Boston, May 10–14, 2010) was awash in analyses of networks obtained by crawling Facebook. Governments want to know about suspicious behavior. Businesses want to know whom to target and whom to leave alone, what products to develop, what infrastructure to plan for, and what charging systems to deploy. Social network analysis has become a huge, high-tech business that involves mathematicians, physicists, computer scientists, and, of course, social scientists. A recent article in *The Economist* [15] quantifies things:

> "The market for such software is booming. By one estimate there are more than 100 programs for network analysis, also known as link analysis or predictive analysis. The raw data used may extend far beyond phone records to encompass information available from private and governmental entities, and internet sources such as Facebook. IBM . . . says its annual sales of such software, now growing at double-digit rates, will exceed $15 billion by 2015. In the past five years IBM has spent more than $11 billion buying makers of network-analysis software."

Mark Rogers, CEO of the London-based Market Sentinel Ltd, which specializes in semantic search and social network analysis, told us that

> "The challenge of big consumer brands is to understand a media landscape fragmented into numerous 'contexts'—semantic and social. A context could be 'people I was at university with' or 'Justin Bieber fans who also like my competitor's product'. Brands have this incredibly rich cluster of groups to understand, enfranchise, co-opt. . . .
>
> "The insights are extraordinary: for example, by using Skyttle to look at co-occurring hashtags on Twitter (if I use a particular hashtag, what other hashtags do I tend to use?) we can determine that people who tweet about how much they love their cats are also likely to tweet about depression. Brands can use the understanding gained from tools like this to drive everything from product design to communications language."

The U.S. National Institutes of Health also take this area very seriously, having launched in 2010 a program in Social Network Analysis and Health, with a view to "encourage basic research that will: generate new theories that can further social network analysis; address fundamental questions about the relationship between social networks and health; and develop methodological and technological innovations to facilitate and extend social network analyses."

The Palo Alto-based company Palantir Technologies, founded in 2004 by a group of PayPal alumni and Stanford computer scientists, offers software that analyses structured, unstructured, relational, temporal, and geospatial data. With no marketing department or sales team, the company employs more than 250 engineers and scientists and has earned contracts with counterterrorism analysts at offices of the FBI and CIA. The company also focuses on data security and fraud prevention; for example, the U.S. government's Recovery Accountability and Transparency Board uses the Palantir government platform to share information with citizens via the web and to detect abuse of the economic stimulus fund.

Of course, a simple network-based view of the world is painted with a very broad brush. On one hand are practical and ethical concerns about gathering and storing data, as in reports [2] of the objections of a former Army colonel alleging that the need for exhaustive network data for counterterrorism purposes in Iraq led to the detainment of locals who were known to be innocent. On the other hand, even with access to all the relevant data, the network view—like any mathematical modelling approach—can represent only a simplified cartoon of a real complex system. As our questions become more probing, the benefits of the pared-down node and edge setting—allowing clean data structures, off-the-shelf algorithms, and clear visualization—start to be outweighed by the loss of information from the oversimplifications. Even allowing for directed or weighted edges does not get around the basic issue: "Nodes" and "interactions" cannot all be treated equally.

David Skillicorn, a computer scientist at Queens University in Ontario, Canada, mentions in his blog (http://skillicorn.wordpress.com/) that work- and leisure-derived connections may need to be treated differently, citing "embarrassing faux pas" resulting from "trying to get people to friend their boss's boss." As an example, he writes,

> "Suppose that you have access to someone's email and you want to work out who are their friends, and who are professional contacts. The structure of email addresses doesn't help much because a friend's work email might be used, and because email addresses tend to be surrogates anyway. Time of day doesn't help much because many people send personal email at work, and many send work emails out of working hours. The content of emails might help, but many organisations have extensive in-house non-work emails (for example, Enron had many emails about fantasy football that circulated only within the company)."

The traditional tools of network science, then, can be only part of the solution. Indeed, the Palantir Technologies website makes it clear that securely integrating heterogeneous data sets is a major challenge:

> "Probably the most central hard problem that we address in trying to enable the analyst is data modeling, the process of figuring out what data types are relevant to a domain, defining what they represent in the world, and deciding how to represent them in the system."

For data sets like those emerging from telecommunications and online social networking, one clear limitation of the network science paradigm seems ripe for mathematizing: Interaction networks do not remain static but, rather, typically evolve over time—"who phoned whom," "who e-mailed whom," and "who Facebook-friended whom" information is inherently dynamic. This issue was raised in August 2010 at the opening workshop of the SAMSI (Statistical and Applied Mathematical Sciences Institute) programme on complex networks. James Moody, a sociologist at Duke University, made the point that if A meets B in the morning and B meets C in the afternoon, an infection can be passed from A to C, but not from C to A. A network summarizing that day's contacts, however, will contain the chain A–B–C, and imposing the usual Susceptible–Infected–Recovered-type model at this level is likely to overestimate the rate at which a disease will spread. Analogously, consider a simple "who phoned whom" scenario, such as that shown in Figure 2, in which the three networks cover successive days. We see that Mary might be able to pass a message to Oscar via the links Mary–Bob–Ramona–Sue–Alf on Day 1 and Alf–Oscar on Day 2, but there is definitely no way for Oscar to get a message to Mary. This asymmetry arises even though each individual network is undirected. Kenth Engø-Monsen, a senior researcher for the Norwegian telecom company Telenor, told us that handling time-dependency may give a key marketing edge:

> "What is marketing in a networked world? It is all about nodes and links! What are the better nodes at spreading your new service or a new product? Over what links is this spreading most probable? The challenge is in identifying the right set of nodes and links that accelerate the diffusion. State-of-the-art network analysis used by telecoms today builds social networks among customers based on traffic data averaged over X weeks or Y months. Taking into account the dynamics of the customer networks is in my mind the next promising thing to consider in network-based marketing techniques."

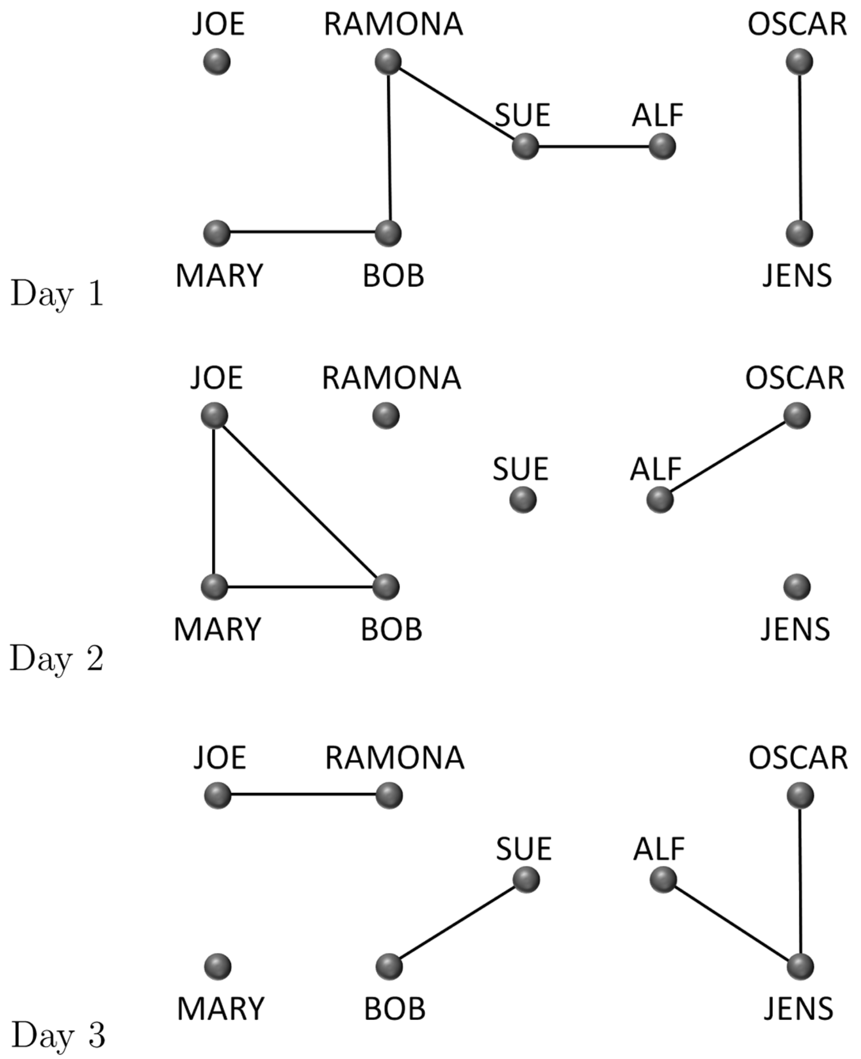

Figure 2. *A simple dynamically varying interaction network.*

One widely studied time-dependent interaction data set in the public domain lists the e-mail activities of 151 Enron employees; see http://www.cs.cmu.edu/~enron/. To produce Figure 3 we first summarized the e-mail interactions, including "to," "cc," and "bcc," over a period of 1138 consecutive days. This leads to a sequence of 1138 symmetric adjacency matrices of dimension 151 × 151, which can also be regarded as a 1138 × 151 × 151 tensor. Figure 3 shows the total number of edges per day. In Figure 4, for the purpose of illustration, we lump the first 1080 days into 12 lots of 30-day "months" and display the monthly adjacency matrices.

Using the notation of [10], for a fixed set of nodes we can consider an ordered sequence of symmetric, binary adjacency matrices, $A^{[k]}$, for $k = 0, 1, 2, \dots, M$, corresponding to an ordered sequence of time points $t_0 \leq t_1 \leq \dots \leq t_M$. Continuing from the examples quoted for Figure 2, a *dynamic walk of length w* from node i_1 to node i_{w+1} could then be defined as a sequence of edges $i_1 \leftrightarrow i_2$, $i_2 \leftrightarrow i_3 \dots i_w \leftrightarrow i_{w+1}$ and a non-decreasing sequence of times $t_{r_1} \leq t_{r_2} \leq \dots \leq t_{r_w}$, such that $(A^{[rm]})_{i_m i_{m+1}} \neq 0$.

This time-dependent context offers a rich variety of alternatives, however. In some circumstances it may be appropriate to allow the use of at most one edge per time point, in which case the constraint on the time sequence would involve strict inequality: $t_{r_1} < t_{r_2} < \dots < t_{r_w}$. At a formal evening ball, for instance, with each dance you have the choice of finding a partner or sitting it out. Alternatively, it may be appropriate to force the use of exactly one edge per time point, so that

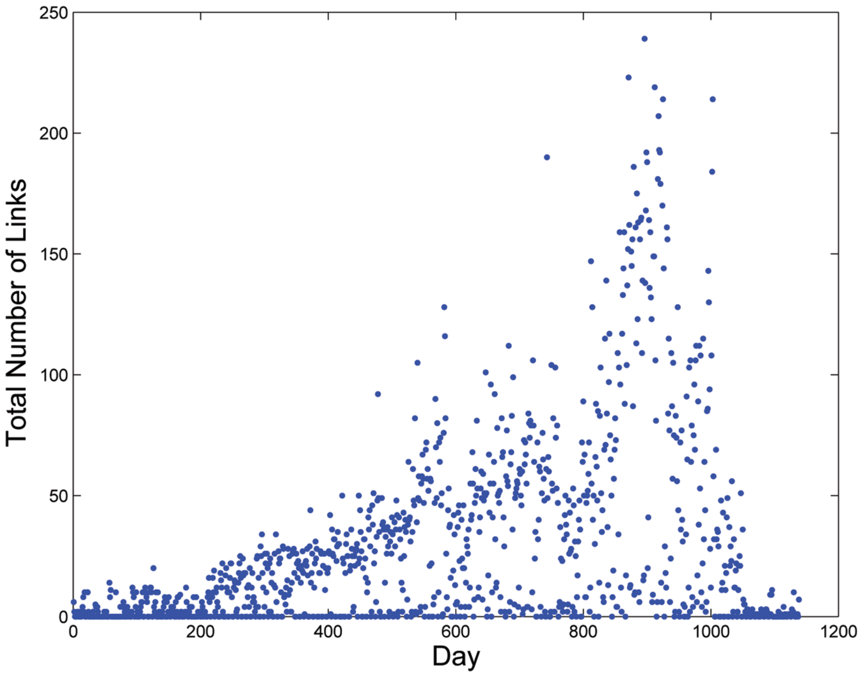

Figure 3. *Total number of links per day in Enron e-mail data.*

$r_{n+1} = r_n + 1$. Suppose, for example, that the school's pet hamster is taken home each day: Yesterday's lucky pupil passes it on to a current friend to take home today.

We can define *dynamic paths* and *dynamic trails* analogously. In the case of a path, we would insist that no node be visited more than once—this concept was developed recently in [14] and used to construct dynamic centrality measures. The issue of discovering well-connected communities in time-varying networks is addressed in [12]. We also note that the extra "time" dimension takes us into the realm of tensors, where data-mining issues are currently an active topic [1]. The dynamic walk concept has been pursued in [10] via the observation that element i,j in the matrix product $A^{[r_1]} A^{[r_2]} \ldots A^{[r_w]}$ counts the number of dynamic walks of length w from node i to node j, where the mth step of the walk takes place at time t_{r_m}. It follows that the matrix product

$$(I - aA^{[0]})^{-1} (I - aA^{[1]})^{-1} \ldots (I - aA^{[M]})^{-1} \qquad (1)$$

does the job of collecting all dynamic walks: Its element i,j gives the overall number of dynamic walks from i to j, where a walk of length w is scaled by a^w. This leads naturally to centrality measures. Furthermore, even though each resolvent $(I - aA^{[r]})^{-1}$ is symmetric, the overall product of resolvents will generally be asymmetric, reflecting the inherent non-commutativity arising from the arrow of time. Results for the Enron data shown in Figures 3 and 4, including quantifications of the best broadcasters and receivers of information, can be found in [10].

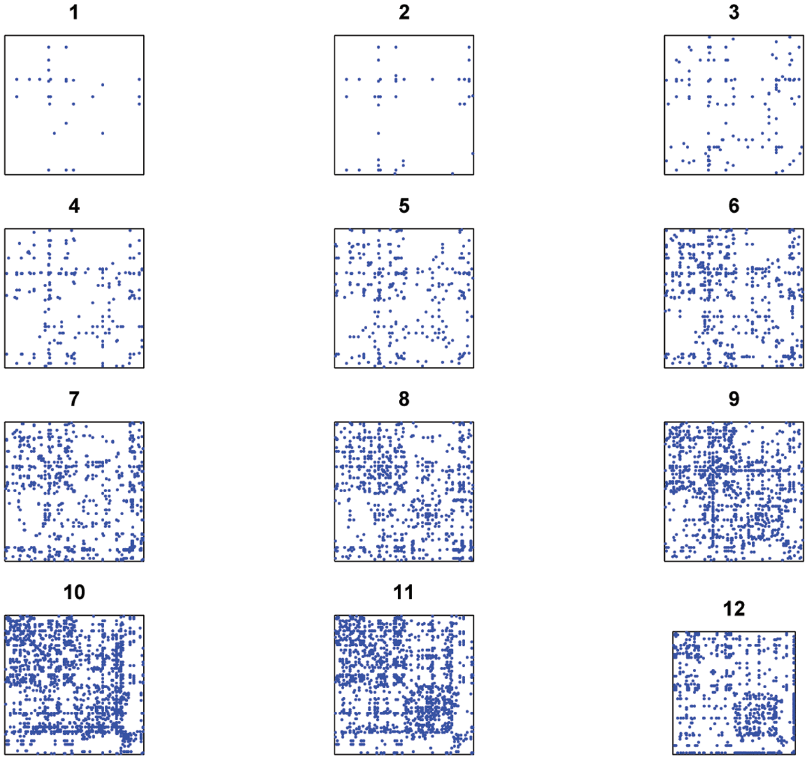

Figure 4. *Evolving adjacency matrices for Enron e-mail data aggregated over 30-day periods.*

To emphasize that these ideas must be fine-tuned to the particular application, we note that in many cases $A^{[k]}$ will itself be an aggregate of activity over a time window, as in the case of daily e-mail communication. Suppose that we refine the time window—to hours, minutes, seconds, ... At some stage the product in (1) would settle at a fixed value. We can see this intuitively from the fact that with at most one link per time period, no further walks can be created through refinement. We could also argue directly from the linear algebra setting of (1) that no new non-identity factors will arise. With such a high sampling rate, it is then tempting to argue that (1) should be replaced with the "at most one link per time window" version

$$(I + aA^{[0]}) (I + aA^{[1]}) \ldots (I + aA^{[M]}),$$

so that simple, inter-window closed walks like $i \mapsto j \mapsto i$ are avoided. As with the bid–ask spread issue in mathematical finance, however, this "high frequency equals high accuracy" argument may be scuppered by noise—in our experience, the order in which e-mails are dealt with does not usually match the order of arrival.

We have focused here on data-driven issues: summarizing, ranking, and extracting key features. Applied mathematicians also strive to understand the mechanisms that govern a system and to express them in terms of a quantitative, predictive model. In the case of evolving networks, what are the laws of motion that regulate the appearance and disappearance of edges? Some basic stochastic models for edge dynamics were proposed and analysed in [9], and the related issues of (a) how to calibrate a model and (b) how to analyse a dynamic process taking place over a dynamic network were addressed. With evolving data of this type, there are some very natural, real-time issues. Given the past behaviour of the network and its current state:

- Is the network operating normally?
- Is a dramatic event (such as a power blackout or computer virus outbreak) about to happen?
- Are there any pockets of suspicious activity?
- Which nodes in the network are currently most vulnerable?
- Is this a good time to start a rumour, or to hear the latest rumour? If so, where?

We began this article (*SIAM News*, January/February, page 1) by arguing that the emerging ideas from the quantitative realm of social network analysis have spawned a fascinating two-way interplay with the mathematical sciences. The recent explosion in the quantity and variety of available data concerning our patterns of behaviour has turned this into a very high-profile pursuit, raising the stakes and opening many new challenges. We finish with a view from Mark Rogers, of Market Sentinel Ltd, who told us that

> "Some of the big industries which will emerge from current developments in semantic and social search are going to be in the marketing/comms area. Conventional market research and product/communications planning are being replaced wholesale. Of course, marketing is only one industry to be impacted in this way, it will be followed by many others."

Acknowledgments

The authors are supported by the UK Engineering and Physical Sciences Research Council Mathematical Sciences programme and the Research Council UK Digital Economy programme.

References

[1] E. Acar, D.M. Dunlavy, and T.G. Kolda, *Link prediction on evolving data using matrix and tensor factorizations*, in LDMTA'09: Proceedings of the ICDM'09 Workshop on Large Scale Data Mining Theory and Applications, IEEE Computer Society Press, December 2009, 262–269.

[2] J. Bohannon, *Counterterrorism's new tool: 'Metanetwork' analysis*, Science, 325 (2009), 409–411.

[3] S.P. Borgatti, *Centrality and network flow*, Social Networks, 27 (2005), 55–71.

[4] E. Estrada, M. Fox, D.J. Higham, and G.-L. Oppo, eds., *Network Science: Complexity in Nature and Technology*, Springer, Berlin, 2010.

[5] E. Estrada and D.J. Higham, *Network properties revealed through matrix functions*, SIAM Rev., 52 (2010), 696–714.

[6] E. Estrada and J.A. Rodríguez-Velázquez, *Subgraph centrality in complex networks*, Phys. Rev. E, 71 (2005), 056103.

[7] L.C. Freeman, *Turning a profit from mathematics: The case of social networks*, J. Math. Sociol., 10 (1984), 343–360.

[8] L.C. Freeman, *Going the wrong way down a one-way street: Centrality in physics and biology*, J. Social Structure, 9 (2008).

[9] P. Grindrod and D.J. Higham, *Evolving graphs: Dynamical models, inverse problems and propagation*, Proc. Roy. Soc., Series A, 466 (2010), 753–770.

[10] P. Grindrod, D.J. Higham, M.C. Parsons, and E. Estrada, *Communicability across evolving networks*, Mathematics and Statistics Research Report 32, University of Strathclyde, 2010.

[11] L. Katz, *A new index derived from sociometric data analysis*, Psychometrika, 18 (1953), 39–43.

[12] P.J. Mucha, T. Richardson, K. Macon, M.A. Porter, and J.-P. Onnela, *Community structure in time-dependent, multiscale, and multiplex networks*, Science, 328 (2010), 876–878.

[13] M.E.J. Newman, *Networks: An Introduction*, Oxford University Press, Oxford, UK, 2010.

[14] J. Tang, M. Musolesi, C. Mascolo, V. Latora, and V. Nicosia, *Analysing information flows and key mediators through temporal centrality metrics*, in SNS 2010: Proceedings of the 3rd Workshop on Social Network Systems, New York, NY, 2010, ACM, 1–6.

[15] Technology Quarterly, *Untangling the social web. Software: From retailing to counterterrorism, the ability to analyse social connections is proving increasingly useful*, The Economist, September (2010).

[16] S. Wassermann and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge, UK, 1994.

[17] D.J. Watts and S.H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, 393 (1998), 440–442.

Desmond J. Higham is a professor in the Department of Mathematics and Statistics, University of Strathclyde, UK. Peter Grindrod is a professor of mathematics in the Department of Mathematics and Statistics at the University of Reading, UK. Ernesto Estrada is a professor in the Departments of Mathematics and Statistics, and Physics, at the University of Strathclyde, UK.