# Visualization and Modeling of Structural Features of a Large Organizational Email Network

Benjamin H. Sims
Statistical Sciences (CCS-6)
Los Alamos National Laboratory
Los Alamos, New Mexico 87545
Email: bsims@lanl.gov

Nikolai Sinitsyn
Physics of Condensed Matter and Complex Systems (T-4)
Los Alamos National Laboratory
Los Alamos, New Mexico 87545
Email: nsinitsyn@lanl.gov

Stephan J. Eidenbenz
Information Sciences (CCS-3)
Los Alamos National Laboratory
Los Alamos, New Mexico 87545
Email: eidenben@lanl.gov

***Abstract*—This paper presents findings from a study of the email network of a large scientific research organization, focusing on methods for visualizing and modeling organizational hierarchies within large, complex network datasets. In the first part of the paper, we find that visualization and interpretation of complex organizational network data is facilitated by integration of network data with information on formal organizational divisions and levels. By aggregating and visualizing email traffic between organizational units at various levels, we derive several insights into how large subdivisions of the organization interact with each other and with outside organizations. In the second part of the paper, we propose a power law model for predicting degree distribution of organizational email traffic based on hierarchical relationships between managers and employees. This model considers the influence of global email announcements sent from managers to all employees under their supervision, and the role support staff play in generating email traffic, acting as agents for managers.**

## I. Introduction

In this paper, we present results of our analysis of a large organizational email dataset, comprising nearly complete email traffic records for Los Alamos National Laboratory (LANL) over a period of three months. Very few organizational communication networks of this scale have been analyzed in the literature. Analyzing such large email datasets from complex organizations poses a number of challenges. First, considerable work is required to parse large quantities of raw data from network logs and convert it into a format suitable for network analysis and visualization. Second, a great deal of care is required to analyze and visualize network data in a way that makes sense of complex formal organizational structures - in our case, 456 organizational units that are connected through diverse organizational hierarchies and management chains. Finally, it can be difficult to sort out the effects of email traffic generated by mass announcements and communications along management chains from the more chaotic, less hierarchical traffic generated by everyday interactions among colleagues.

This paper addresses these complexities in two ways. First, we demonstrate methods for understanding large-scale structural relationships between organizational units by using carefully thought-out visualization strategies and basic graph statistics. Second, we propose a power law model for predicting the degree distribution of email traffic for nodes of large degree that engage in mass emails along hierarchical lines of communication. This likely characterizes a significant portion of email traffic from managers (and their agents) to employees under their supervision.

## II. Analysis of Organizational Structure

While many analysts have examined ways of extracting structural features from corporate email exchange networks, they have typically focused at the level of email exchanges between individuals (albeit sometimes large numbers of individuals), bringing little or no information about formal organizational structures into their analysis [1], [2], [3]. Aggregating relationships based on formal organizational structures offers another important level of insight, which can be particularly useful for managers and analysts interested in interactions among business units, capabilities, or functions rather than individuals. Automatically collected email data has significant advantages for capturing interactions among organizational units: although email does not capture all relevant interactions, it provides comprehensive coverage across the entire organization without the overhead involved in large-scale survey-based studies. In order to locate individuals within organizational structures, we used organizational telephone directory data to associate email addresses with low-level organizational units, and organization charts to generate mappings of these units to higher-level ones.

### *A. Structural relationships between elements of the organization*

Our analysis of structural relationships within LANL focuses on two broad, cross-cutting distinctions: program vs. line organizations, and technical research and development functions vs. operations functions (safety, physical plant, etc.)

LANL is a hybrid matrix management organization. In a true matrix organization, individuals are assigned to line management units that are based entirely on capabilities - for example, a Statistical Sciences group or a Physics group - but their work is directed by project managers, who recruit

employees from line organizations to work on projects. This model became popular in the aerospace industry with the rise of program management in the 1950s, and was in part influenced by the organizational structure of the Manhattan Project [4]. At LANL today, line and program organizations play less distinct roles. The base-level line units that house most employees are called groups, which may be built around programs or capabilities. In our analysis, we draw a distinction between groups and higher-level line management organizations, which aren't directly involved in technical or operations work. Program organizations play a variety of coordinating roles among groups, management, and outside organizations, and sometimes conduct technical or operations work as well. Despite this flexible definition, our analysis reveals that technical program organizations occupy a very well-defined structural space within the organization as a whole.

Fig. 1 shows email traffic between all organizational units at LANL, laid out using a force-vector algorithm. The units are colored according to the classification described above, and their sizes represent betweenness centrality. There are some visible patterns in this layout. First, a number of operations groups have the highest centrality, probably because they provide services to most of the other organizational units at the laboratory. Ranking the nodes by betweenness centrality confirms this: 17 of the top 20 nodes are operations organizations. In addition, operations units and technical units occupy distinct portions of the graph; this indicates that there is generally more interaction within these categories than between them. The highly central operations groups appear to play a bridging role between the two categories. Administration units appear to be somewhat more closely associated with technical units than operations units, although this is difficult to state with certainty.

Some of the ambiguities in interpretation can be clarified by grouping all units in a given category into a single node, resulting in the 7-node graph shown in Fig. 2. This view, which uses a simple circular layout, reveals that there is a large amount of email traffic (in both directions) on the technical side of the organization along the path Administration - Management - Program - Group, and relatively little traffic between these entities along any other path. This suggests that program organizations, rather than representing an independent chain of command as in a true matrix organization, instead play an intermediary role between technical groups and technical management. The operations side of the organization does not display this pattern, indicating that relationships between groups, programs, and management are more fluid there.

Another way of understanding the roles different types of organizational units play is in terms of their relationships with outside entities. Fig. 3 plots the number of emails each type of organization sends and receives to/from commercial vs. non-commercial domains. This indicates that all types of operational units communicate significantly more with commercial entities, which is probably driven by relationships with suppliers and contractors. Technical groups, technical management, and administration communicate about equally with commercial and non-commercial domains. The outlier here is technical programs, which are much more highly connected to non-commercial domains, particularly .gov addresses. This further expands on the role of technical programs, suggesting that they are a nexus for coordination of technical work both internally, among line management organizations, and externally, between LANL and outside funding agencies. This is a potentially important finding, with implications for how program organizations should be supported and managed.

### *B. Structural relationships within organizational units*

Email network maps can also be used to visualize relations among members of an organizational unit. Figures 4 and 5 show email networks that were obtained from email exchange records among the members of two LANL groups. We intentionally chose groups that do similar work (theoretical research). In the smaller group in Fig. 4, the two nodes with highest betweenness centrality are group managers, and the third is technical support staff. Thus, the group has a relatively unified hierarchical structure with management and support staff at the center. In the larger group, managers were still among the most central nodes, but many other nodes had similar betweenness centrality (Fig 5). These include administrative assistants, seminar organizers, and several project leaders. This indicates a flatter, less centralized organizational structure. This group divides roughly equally into two dominant communities. This divide seems to reflect the fact that the group was recently formed by merging two previously existing groups.

## III. Node connectivity distribution as a function of organizational hierarchy

Several network types, including biological metabolic networks [7], the World Wide Web, and actor networks [8], are conjectured to have power law distributions of node connectivity. In the case of metabolic networks, the interpretation of scale free behavior is complicated by the lack of complete knowledge and relatively small sizes ($\sim 10^3$ nodes) of such networks, while the mechanisms of self-similarity in many large social networks are still the subject of debate. However, organizational hierarchy has been shown to generate degree distributions for contacts between individuals that follow power laws [9].

Managers prefer to use email to communicate with subordinates in many different communication contexts [10]. We propose that node connectivity patterns in the email networks of large, formal organizations are driven, in part, by management hierarchy and specific patterns of email use by managers, in particular the mass broadcast of email announcements. Based on this observation, we develop a scale-free behavioral model that takes into account features specific to email communications in organizations. In this model, the self-similarity of the connectivity distribution of the email network is a consequence of the static self-similarity of the management structure, rather than resulting from a dynamic process, such

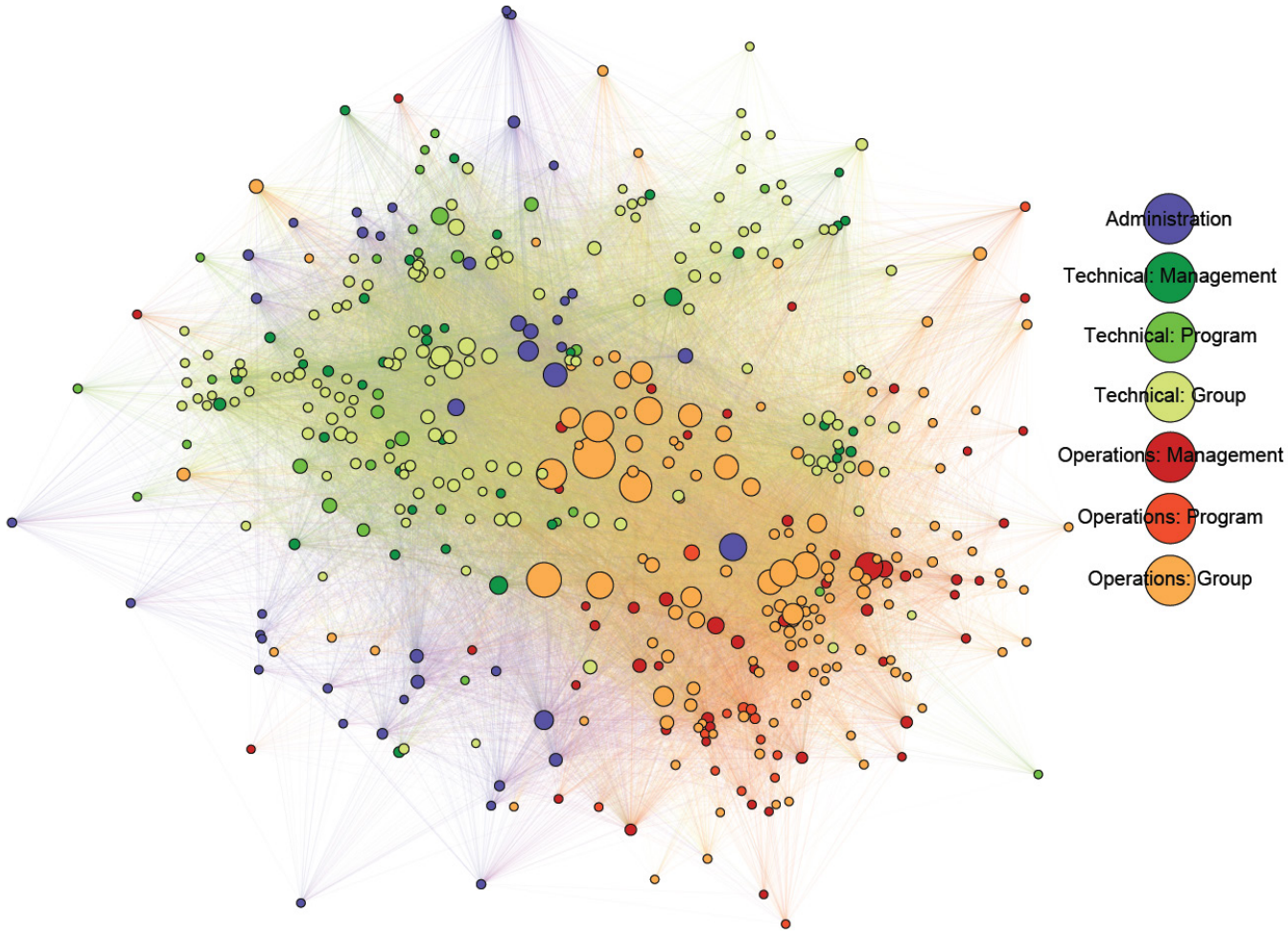


Fig. 1. Email traffic between organizational units at LANL, using a force-vector layout. Node size represents betweenness centrality. Edge color is a mix of the colors of the connected nodes. Although individual edges are difficult to discern at this scale, the overall color field reflects the type of units that are most connected in a given region.

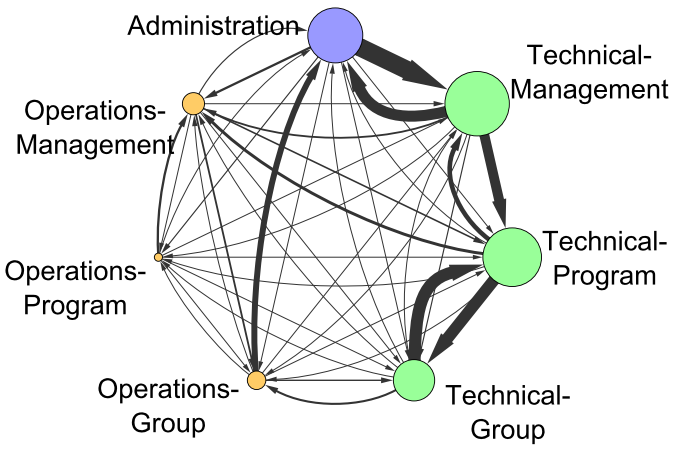


Fig. 2. Email traffic between organization types at LANL. Node diameter represents total degree (i.e. total number of incoming and outgoing emails) of the node; edge width represents email volume in the direction indicated.

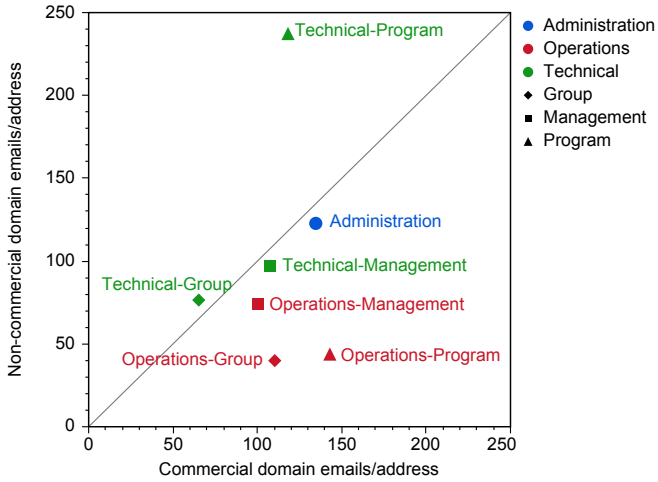


Fig. 3. Total emails to/from commercial (.com, .net, .info) vs. non-commercial (.gov, .edu, .mil, etc.) domains, by organization type.

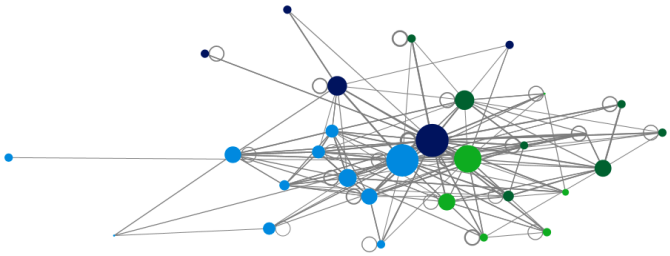

Fig. 4. Email network for 2 week period in smaller group. Size of a node is proportional to logarithm of its betweenness centrality. Nodes with different colors correspond to different communities that were identified by application of the Girvan-Newman algorithm to the group's email network [5], [6]. Link widths are proportional to the logarithm of the number of emails exchanged along these links. The network was visualized by assigning repulsion forces among nodes and spring constants proportional to the link weights, and then finding an equilibrium state.

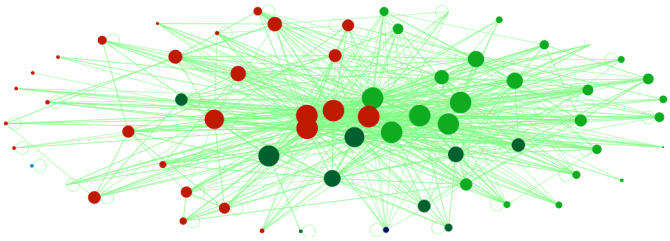

Fig. 5. Email network for 2 week period in larger group.

as preferential attachment [11] or optimization strategies [12]. More specifically, self-similarity is due to the ability of a manager to continuously and directly communicate only with a relatively small number of people, while communications with other employees have to be conveyed in the form of broad announcements.

Suppose that the top manager in an organization sends emails to all employees from time to time. This manager must correspond to the node in the email network that has highest connectivity $N$. Suppose that the top manager also talks directly (in person) to $l$ managers that are only one step lower in the director's hierarchy (let's call them 1st level managers). Each of those 1st level managers, presumably, control their own subdivisions in the organization. Assuming roughly equal spans of managerial control, we can expect that, typically, one 1st level manager sends emails to $N/l$ people. In reality, each manager also has a support team, such as assistants, administrators, technicians, etc. who also may send announcements to the whole subdivision. Let us introduce a coefficient $a$ which says how many support team employees are involved in sending global email announcements in the division on the same scale as their manager. We can then conclude that at the 1st level from the top there are $al$ persons who send emails to $N/l$ employees at a lower level.

Each 1st level manager controls $l$ 2nd level ones and we can iterate our arguments, leading to the conclusion that there should be $(al)^2$ managers on the 2nd level who should be connected to $N/(l^2)$ people in their corresponding subdivisions. Continuing these arguments to the lower levels of the hierarchy, we find that, at a given level $x$, there should be $(al)^x$ managers (or their proxies) who write email announcements to $N/(l^x)$ people in their subdivision.

Consider a plot that shows the number of nodes $n$ vs. the weight of those nodes, i.e. their outdegree $w$. Considering previous arguments, we find that the weight $w = N/(l^x)$ should correspond to $n = (al)^x$ nodes. Excluding the variable $x$, we find

$$\log(n) = \frac{\log(al)}{\log(l)} \left(\log(N) - \log(w)\right), \qquad (1)$$

where $\log$ is the natural logarithm.

Eq. (1) shows that the distribution of connectivity, $n(w)$, in a hierarchical organizational email network should generally be a power law with exponent $\frac{\log(al)}{\log(l)} > 1$. Obviously, at some level $x$, this hierarchy should terminate around the point at which $(al)^x = N/(l^x)$, because the number of managers should not normally exceed the number of employees. Hence the power law (1) is expected to hold only for nodes with heavy weights, e.g. $n > 50$, i.e. for nodes that send announcement-like one-to-many communications, and at lower $n$ this model predicts a transition to some different pattern of degree distribution. At this level, it is likely that non-hierarchical communication patterns begin to dominate in any case.

In order to compare this model to actual network data, we analyzed the statistics of node connectivity in email records at LANL during a two-week time interval (Fig. 6). We removed nodes not in the domain $lanl.gov$ and cleaned the database of various automatically generated messages, such as bouncing emails that do not find their target domain. However, we kept domains that do not correspond to specific employees, such as emails sent from software support services. Our remaining network consisted of $N \approx 32000$ nodes, which is still about three times the number of employees at LANL. This is partially attributed to the fact that we did not exclude domains that are not attached to specific people, and also the fact that a significant fraction of employees have more than one email address for various practical reasons.

Numerical analysis, in principle, should allow us to obtain information about parameters $l$, $x$ and $a$, from which one can make some very coarse-grained conclusions about the structure of the organization. Such an analysis should, of course, always be applied with a certain degree of skepticism due to potential issues with data quality, the simplicity of the model, and logarithmic dependence of the power law on some of these parameters [13]. We found that our data for $w > 40$ could be well fitted by $\log(n) \approx 14.0 - 2.47\log(w)$ (Fig. 7). If, e.g., we assume $l = 4$, then $a \approx 7$, i.e. each manager has the support of typically $a - 1 = 6$ people, who help her post various announcements to her domain of control. The power law should terminate at the level of hierarchy $x$ given by $(al)^x = N/(l^x)$, which corresponds to $x \approx 3$, i.e. the email network data suggest that there are typically $x = 3$ managers of different ranks between the working employee and the top manager of the organization. The typical number

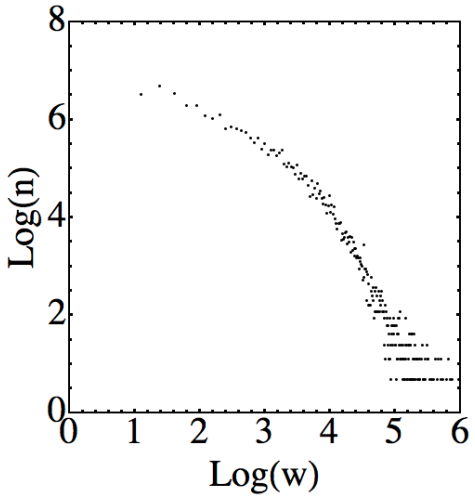


Fig. 6. LogLog plot of the distribution of the number of nodes $n$ having the number of out-going links $w$.

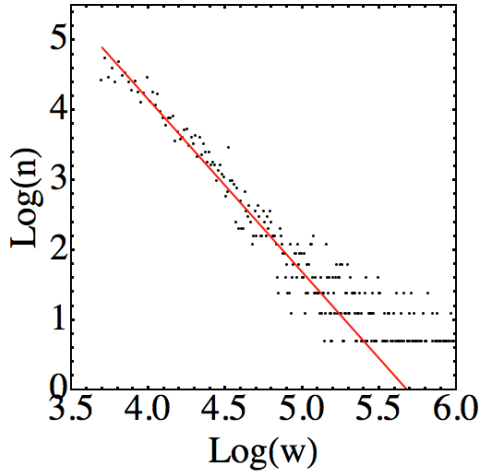


Fig. 7. Zoom of Figure 6 for $w > 40$. Red line is a linear fit corresponding to $\log(n) \approx 14.0 - 2.47\log(w)$.

of email domains to which the lowest rank manager sends announcements is $w_{\min} \approx N/l^x \approx 48$. This should also be the degree of the nodes at which the power law (1) should be no longer justified. Indeed, we find the breakdown of the power law (1) at $w < 40$. This estimate also predicts that a typical working employee receives emails from $(x+1)a = 28$ managers or their support teams.

Comparing these results to the actual organizational structure of the organization is very difficult due to the large excess of email addresses over the number of actual employees, and the lack of empirical data on many of the model parameters. Keeping in mind these difficulties, the estimated model parameters seem to be generally consistent with the actual organizational structure. In reality, LANL has 5 possible layers of line management between an employee and the laboratory director, but this is complicated by the facts that the lowest layer is often not used, and some employees work for organizations that report directly to a higher-level manager. So the estimate of $x \approx 3$ given above might be consistent with the actual organization structure. The average group size at LANL is difficult to determine quantitatively from available data, but appears to be generally in the 20-40 person range, which is somewhat lower than the number of domains (48) to which the lowest-level manager sends emails based on model estimates. Again, although these results might suggest possible conclusions about the accuracy of the model, we do not currently have data of sufficient quality to make a rigorous comparison between model estimates and real-world organizational structure in this case.

## IV. Conclusion

Visualizing and modeling email traffic in complex organizations remains a challenging problem. Visualizing email data in terms of formal organizational units reduces complexity and provides results that are more intelligible to organization members and analysts interested in understanding organizational structure at a macro level. For predicting the degree distribution of high-degree nodes in an organization, we find that it is useful to take into account both organizational hierarchy and email-specific behavior (in particular, the use of mass emails within line management chains). These findings suggest that considering information about formal organizational structures alongside email network data can provide significant new insights into the functioning of large, complex organizations.

## References

[1] J. Diesner, T. L. Frantz, and K. M. Carley, "Communication networks from the Enron email corpus 'It's always about the people. Enron is no different.' " *Computational and Mathematical Organizational Theory*, vol. 11, no. 3, pp. 201-228, Oct. 2005.

[2] A. Chapanond, M. S. Krishnamoorthy, and B. Yener, "Graph theoretic and spectral analysis of Enron email data," *Computational and Mathematical Organizational Theory*, vol. 11, no. 3, pp. 265-281, Oct. 2005.

[3] T. Karagiannis and M. Vojnovic. (2008, May). *Email Information Flow in Large-Scale Enterprises* [Online]. Available: http://research.microsoft.com/pubs/70586/tr-2008-76.pdf.

[4] G. E. Bugos, "Programming the American aerospace industry, 1954-1964: The business structures of technical transactions," *Business and Economic History*, vol. 22, no. 1, pp. 210-222, Fall 1993.

[5] D. L. Hansen, B. Shneiderman, and M. A. Smith, *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. Burlington, MA: Elsevier, 2011.

[6] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *PNAS*, vol. 99, no. 2, pp. 7821-7826, Apr. 2002.

[7] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabasi, "Hierarchical Organization of Modularity in Metabolic Networks," *Science* vol. 297 no. 5586 pp. 1551-1555, Aug. 2002.

[8] E. Ravasz and A.-L. Barabasi, "Hierarchical organization in complex networks," *Phys. Rev. E*, vol. 67, no. 2, 026112, Feb. 2003.

[9] A.-L. Barabasi, E. Ravasza, and T. Vicsek, "Deterministic scale-free networks," *Physica A*, vol. 299, no. 3-4, pp. 559-564, Oct. 2001.

[10] M. L. Markus, "Electronic Mail as the Medium of Managerial Choice," *Organization Science*, vol. 5, no. 4, pp. 502-527.

[11] M. Mitzenmacher, "A brief history of generative models for power-law and lognormal distributions," *Internet Mathematics*, vol. 1, p. 226, 2004.

[12] F. Papadopoulos, M. Kitsak, M. A. Serrano, M. Boguna, and D. Krioukov, "Popularity versus similarity in growing networks," *Nature* vol. 489, p. 537, 2012.

[13] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661-703, 2009.