# Understanding the Structure of Links in Networks

## CS 322 Final Project Report

Tal Rusak
Computer Science Department
Stanford University
Stanford, California, USA 94305
tal@cs.stanford.edu

## ABSTRACT

We observe non-trivial temporal variation and correlation in the usage patterns in a wide range of social network links. In particular, we note that the usage of links is highly variable and temporally correlated (bursty) over a large number of time scales. This is in contrast to a naïve analysis that might assume that use of Internet-enabled social networks smoothes as the aggregation scale increases.

To quantify these variations, we introduce several statistical metrics and evaluate them on real data collected from a number of diverse social networks. In addition, we show how to interpret these metrics in terms of scales that naturally apply in human dynamics—days, weeks and months—and we find commonalities among different networks in their behavior at these natural time scales. Finally, we define statistical metrics that can be summarized as vectors and used to label nodes or edges in networks for studies of correlation or information flow.

## Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Modeling techniques; G.3 [**Probability and Statistics**]: Time series analysis

## General Terms

Measurement, Performance, Experimentation

## Keywords

Social networks, Burstiness, Time Scale Analysis

## 1. INTRODUCTION

In this paper, we study the *burstiness* of activity over electronically-mediated social network systems and, in particular, over network links. In bursty links, traffic patterns tend to be correlated in time, with busier periods followed by more busy periods with a high likelihood and lightly loaded periods followed by more lightly loaded periods with a high likelihood. We investigate this property in social networks such as several e-mail networks and the memetracker network. As a baseline, we investigate the aggregate communications activity in these networks and draw conclusions about the "bandwidth" of the communications link—the amount of traffic that the link sustains as time passes. We also consider the behavior of individual network links and define vectors representing several statistical properties that can be used for clustering and correlation analysis.

The rest of the paper is organized as follows: in Section 2, we review related work; in Section 3, we summarize the network data used for this study; in Section 4, we define and demonstrate the statistical metrics; in Section 5, we discuss the lessons and applications provided by these techniques; in Section 6, we suggest a model for bursty links; and in Section 7, we provide concluding remarks and suggest directions for future work.

## 2. BACKGROUND AND RELATED WORK

One of the first studies of social networks can be traced to Leland et. al's seminal study of the variation of Ethernet traffic in time [6]. Although not viewed at the time as a social network, the work examined aggregate Ethernet network traffic over a large research institution, the former Belcore labs. Since most Ethernet traffic is generated by human users, the variation in traffic generally corresponds to human behavior. This study concluded that network traffic, in aggregate, was *self-similar*—it varied consistently over many time scales.

There is an elegant social story that helps explain this observation: the network is used more heavily during work weeks than on holidays; it is used more on weekdays than on weekends; it is used more during work hours than at night; and it is used more at 9 AM and 3 PM than at 6 AM and noon. Certain events, such as news events, deadlines for large projects, etc., cause many people to utilize communications links at once. The heavier (and more lightly loaded) periods are likely to be followed by additional heavy (or light) periods, suggesting burstiness. Furthermore, a similar story is true at hugely divergent time scales.

Subsequent studies of social network links have generally not considered time series analysis of aggregate or individual behavior. Instead, more general and time-independent conclusions have been reached. For example, Kossinets and Watts [5] consider a large e-mail network from a university, and make conclusions about cyclic, focal, and triadic closure and the clustering coefficient. Several papers consider the distribution of interarrival times in e-mail networks [3,

8, 14]. Although there is a consensus that this distribution is heavy tailed, this may due to natural cycles in human activities [8] as opposed to a fundamental characteristic of the e-mail channel.

The characteristics of channel variation over multiple time scales have been studied in the context of low-power wireless networks. In the context of these studies [10, 11], pervasive burstiness and self-similarity were observed. We posit that similar characteristics exist in social networks and investigate them further in this paper.

# 3. NETWORK DATASETS

We consider three diverse social networks in this study: the Meme-Tracker dataset [7], the Enron e-mail dataset [4], and an e-mail dataset from a large research institution.

**Meme-Tracker Dataset.** We consider the time series of the number of memes posted to more than 1.6 million blogs and 20,000 sites indexed by Google News. The data set under consideration includes all articles from a 184-day period in late 2008 and early 2009 covering the US presidential election. We capture the number of memes at a one-second and one-minute granularities.

**Enron E-mail Dataset.** We also consider about a half million e-mails from 150 Enron Corp. executives collected and released in court cases following the collapse of the company. We do minimal cleansing of the data by removing any message time-stamped before 1990 or after the end of 2002. We consider all messages in this dataset, including both those sent to and received by Enron employees.

**Research Institution E-mail.** Finally, we study e-mail from a large research institution that tracks 2,397,402 messages over a period of about 1.5 years. There are a total of 225,409 senders in the full dataset, but only about 2,500 addresses have a domain corresponding to that of the institution. We consider both the aggregate e-mail traffic and just those messages sent by members of the institution.

This follows from the observation that sent messages may be a more reliable indicator of e-mail communication than received messages—sent messages do not include any spam that remains undetected, for example [12].

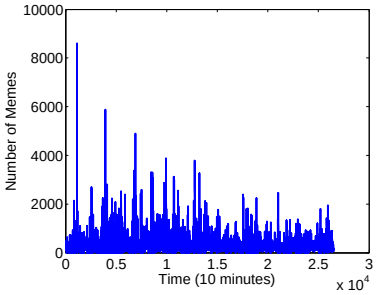
# 4. STATISTICAL METRICS

To study the burstiness and time-varying characteristics of social network links over a wide range of time scales, we considered a number of statistical metrics. We detail these metrics and provide examples.

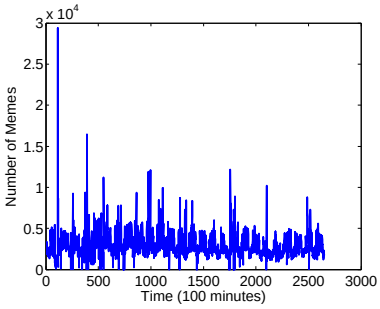
## 4.1 Lens Plot

A *lens plot* is a set of diagrams illustrating the volume of a certain traffic—number of packets received, number of e-mails sent and/or received, number of news articles posted, etc.—aggregated over different time scales. An example for the aggregate Research Institution e-mail dataset was provided in the project milestone, and in Figure 1 we consider the lens plot for the aggregate meme tracker data. Each point on each plot represents the number of items received in the time unit corresponding to this plot.

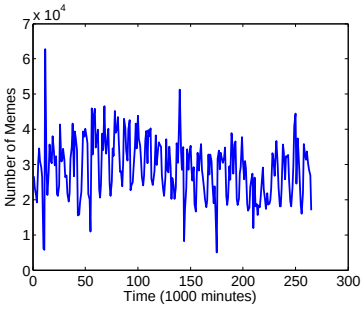
Formally, a lens plot has three parameters: $minScale$, the minimum time scale represented (with temporal dimensions), $maxScale$, the maximum time scale represented (also with temporal dimensions), and $scaling$, the unitless multiplicative factor used to "step" from the minimum time scale



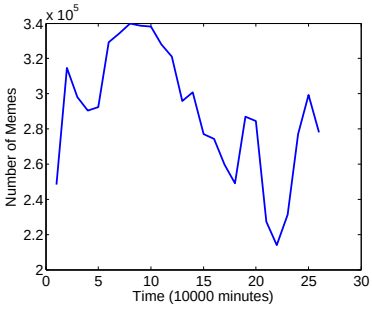

(a) 10 minute lens.



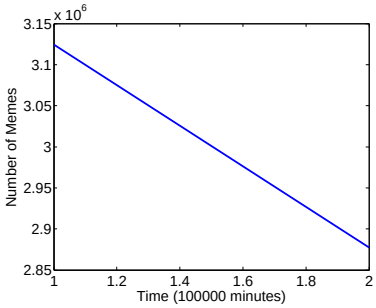

(b) 100 minute lens, corresponding to about 1.7 hours.




(c) 1,000 minute lens, corresponding to about 17 hours.




(d) 10,000 minute lens, corresponding to about 6.9 days.




(e) 100,000 minute lens, corresponding to about 2.3 months.

**Figure 1: Lens plot for Meme-Tracker data. The number of memes parsed is aggregated (summed) over a variety of different time scales. Each point on each plot is the number of memes parsed on the web per time unit. Here, $minScale = 10$ minutes, $maxScale = 100,000$ minutes, and $scaling = 10$.**

to the maximum time scale.

An advantage of the lens plot is that it allows us to quickly and heuristically identify many statistical elements of the data. First, we can easily identify burstiness or smoothing in the data. In a naïve Poisson model, one would expect significant variations at the lowest time scales, as we see in Figure 1(a). However, one would also expect the aggregate traffic to smooth out as the aggregation time scale gets larger [6]. We see in Figures 1(d) and 1(e) that this is not the case in aggregate social network traffic. We also observed similar results in the Enron e-mail dataset and the Research Institution dataset with both all senders and exclusively the institution's senders. Finally, we considered the lens plots of the activity of individual senders in addition to aggregate, network-level metrics. Similar results have been observed on traffic and channel behavior of other networks [6, 10].

Another useful characteristic that we can easily observe from lens plots is that, when examined closely, they show nonstationarities that are inherent in human-generated content due to the natural Circadian and weekly cycles, for example [8]. For example, a nonstationarity is seen in Figure 1(c). Although nonstationarities are inherent and can be easily confused with burstiness, we believe that the burstiness studied here is a separate and important phenomenon.

## 4.2 Normalized Lens Plot

Although it provides an intuitive picture of a link's behavior, one challenge that the lens plot introduces for quantitative statistical analysis is the variable units on the vertical axis. To address this problem, we introduce the *normalized lens plot*. Instead of plotting the absolute volume of messages over different time scales, the normalized lens plot plots this value as a fraction of the maximum volume of messages over some fixed window of the values at the same time scale.

Formally, we use the same parameters as the lens plot—*minScale*, *maxScale*, and *scaling*. We also introduce a fourth parameter for the normalized lens plot, *maxRange*, defining the number of data points, at the current scale, over which the maximum is taken for normalization.

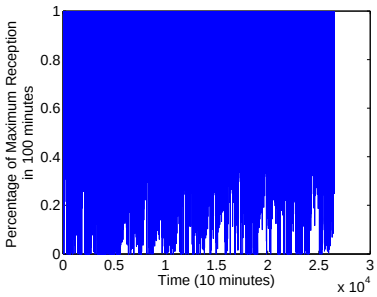
An example of a normalized lens plot is shown in Figure 2, for the same data illustrated in Figure 1 for the Meme-Tracker network.

## 4.3 Standard Deviation Vector

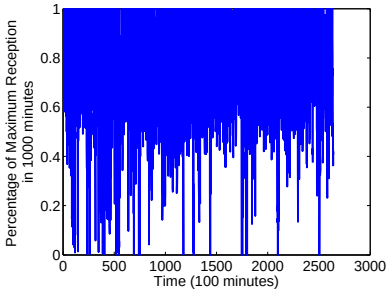
As a first-order statistical metric of burstiness over a large number of time scales, we define the *standard deviation vector*, a vector of standard deviations over the time series represented by the normalized lens plots, with the first element of the vector corresponding to the smallest time scale. This is made possible since the unit on the vertical axis is now normalized. For example, for the plot in Figure 2, the standard deviation vector $\mathbf{s}$ is

$$\mathbf{s} = \begin{pmatrix} 0.2934 \\ 0.2338 \\ 0.1958 \\ 0.0923 \end{pmatrix}.$$

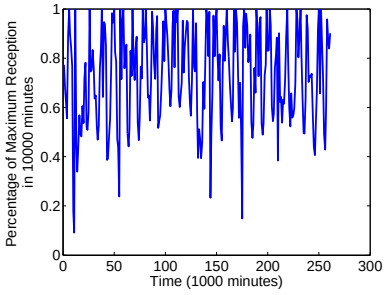
If the values in the standard deviation vector drop sharply, this strongly suggests a smoothing time series, such as a Poisson time series. As we see in this example and in several others we investigated, the standard deviation remains high (almost 10% of the normalized values at the highest



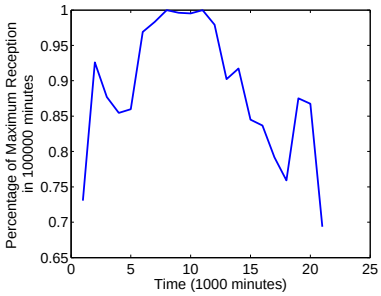

(a) 10 minute normalized lens.




(b) 100 minute normalized lens, corresponding to about 1.7 hours.




(c) 1,000 minute normalized lens, corresponding to about 17 hours.




(d) 10,000 minute normalized lens, corresponding to about 6.9 days.

**Figure 2: Normalized lens plot for Meme-Tracker data. The number of memes parsed is aggregated (summed) over a variety of different time scales and then divided by the highest value in a fixed window of size 10 for the present time scale. Here, $minScale = 10$ minutes, $maxScale = 10,000$ minutes, $scaling = 10$, and $maxRange = 10$.**

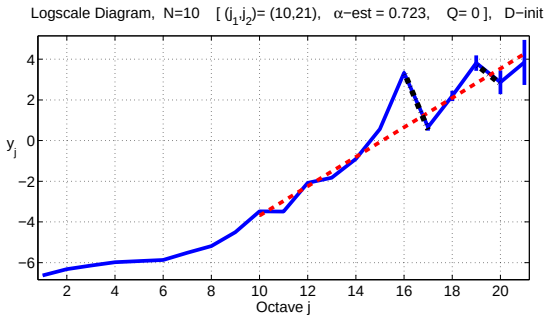



**Figure 3: Logscale diagram for the Research Institution e-mail dataset time series, counting all messages sent by any member of the institution at the resolution of seconds. The solid blue line are the computed logscale values $y_j$, the dotted red line is a fit over part of the data ($\alpha = 0.723$), including the largest time scales, the black dash-dot lines are drops in $y_j$ that may correlate to time scales inherent to human behavior.**

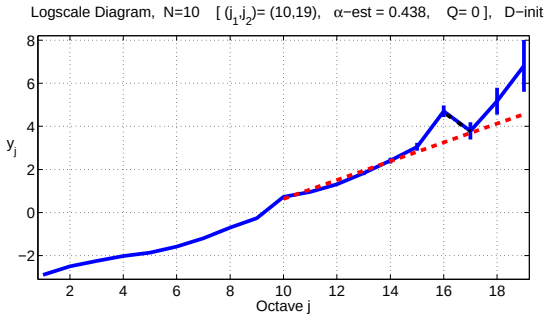



**Figure 4: Logscale diagram for the Meme-Tracker dataset time series, counting all memes posted at the resolution of seconds. Same conventions as in Figure 3, with $\alpha = 0.438$ for the fit shown.**

time scale), strongly suggesting burstiness in the time series. Again, this metric does not require network-level aggregation of links and can be used on individual links or smaller groups of links.

## 4.4 Wavelet Analysis of Scaling

One traditional method to investigate burstiness in network data, termed *scaling* in the statistical literature [1], uses the wavelet-based *logscale diagram*. Figures 3, 4, and 5 provide examples for the aggregate datasets studied in this paper. Formally, on the horizontal axis is an octave, or a base-two time scale of aggregation. Octave 1 corresponds to $2^1 = 1$ time units, octave 2 corresponds to $2^2 = 4$ time units, and so forth. On the vertical axis we plot $y_j$, the logarithm of an unbiased estimator for the variance of the detail wavelet transform at scale $j$ [1]. The slope of the asymptotic domain of the logscale diagram (i.e., the region including the largest time scale) is called $\alpha$ can be used to make conclusions regarding the type of scaling present in the data.

Although a logscale diagram can always be used to provide

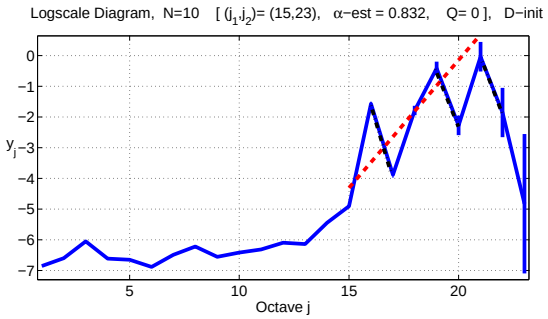



**Figure 5: Logscale diagram for the Enron e-mail dataset time series, counting all memes posted at the resolution of seconds. Same conventions as in Figure 3, with $\alpha = 0.832$ for the fit shown.**

a value for $\alpha$, we must be careful in making a conclusion about what scaling properties are present in the input time series. For example, if the logscale diagram is best fit with two separate lines over different sets of octaves, then the data exhibits *biscaling* [1]. As another example, a logscale diagram for data generated uniformly at random will have a slope of $\alpha = 0$ over most time scales.

While the logscale diagram method is considered to be statistically robust against non-stationarities [1], such phenomena in the data can still affect the results; this is apparently what we see in the examples for the present data.

Finally, we note that we can take a logscale diagram of any time series, whether raw data, aggregated data, or normalized data, such as the normalized lens plot at a certain scale. Furthermore, we can also consider the logscale diagram as a vector of a small number of values that can capture rich details about the behavior of a link in a compact fashion.

# 5. DISCUSSION

The time series analysis conducted leads to several conclusions regarding social networks and suggests methods for analysis.

## 5.1 Natural Time Scales in Logscale Diagrams

In Figures 3, 4, and 5, we see interesting and recognizable behavior arising despite the significant transformation of data performed to derive these plots. In particular, for all three social networks studied, we see major declines in $y_j$ at the octaves $j = 16$, 19, and 21 where these scales are visible. In fact, these are essentially the only major declines in $y_j$ except for octave $j = 22$ for Figure 5, which has a very large error bar since it was computed using few datapoints as the time scale is very large.

The octave $j = 16$ corresponds to $2^{16}$ seconds $= 0.75$ days. This is as close to the "one-day" period as can be seen on a base-two time scale. Similarly, $j = 19$ corresponds to $2^{19}$ seconds $= 6.07$ days, the closest period to "one-week" as can be seen on a base-two time scale. Finally, $j = 21$ corresponds to $2^{21}$ seconds $= 24.3$ days, the closest period to "one-month" as can be seen on a base-two time scale.

There are clear stories for non-stationarities at the day- and week-long scales. Thus, this may be a potential explanation for the drop in the variance of the detail wavelet coefficient which is represented by $y_j$. At the month-long

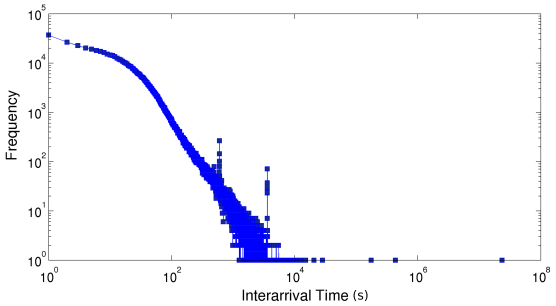



**Figure 6: Probability distribution of e-mail interarrival time over the aggregate Research Institution dataset.**

time scale, the story to support non-stationarities is not as clear; perhaps this has to do with holidays that occur approximately monthly. However, it is interesting and informative to see these natural human time partitions arise out of the analysis of time series data of human behavior without explicitly considering the periodic nature of these time series.

## 5.2 Distribution of Interarrival Time in E-mail

In Figure 6, we plot the distribution of the interarrival time over the aggregate research institution e-mail network. This plot appears to have a heavy tail with some noisy components, as suggested by the previous work [3, 8, 14]. The exact distribution of this curve has been under some dispute, originally believed to be a power law (potentially with exponential cutoff) with a slope of $-1$ for e-mail and $-3/2$ for postal mail [14]. Further evidence has suggested that the distribution is actually lognormal, with nonstationarity elements of human behavior (in particular daily and weekly cycles) [8]. We fit a power law to our overall distribution with a slope of $-1.93$ using Clauset, Shalizi and Newman's estimator [2]. However, it is more likely that the distribution illustrated in Figure 6 fits a double Pareto lognormal distribution [9, 13].

## 5.3 Applications to Local Analysis, Network Structure, and Information Propagation

The statistical methods developed provide an elegant framework to compactly summarize the availability characteristics of a link—in particular, the standard deviation vector and logscale diagram, which can also be considered a vector, provide a summary to thousands of numbers in the lens plots.

A basic capability that these measures provide network researchers is the ability to categorize links as *bursty*, as the previous examples in this paper demonstrate, or *stable*, such as a link whose usage does not change substantially over time. An example of a stable link is one that is used in an purely oscillatory fashion corresponding to weekly, daily, and yearly patterns in human behavior. Another example is a link that is never used.

We can generate these vectors, and perhaps the dichotomous conclusion of burstiness versus stability, for activity originating from nodes or along particular links, thus allowing for clustering and correlation analysis. Figure 7 gives an example of how we can label a graph using these measures.

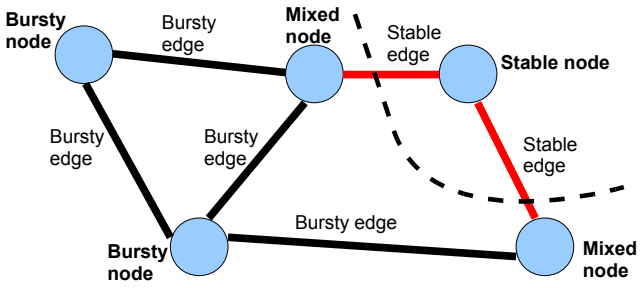



**Figure 7: Example of application of statistical measures to network-level reasoning. In this figure, each edge is labeled in regular type with either a "bursty" or "stable" characteristic, based on the standard deviation vector or logscale diagram. In turn, each node is labeled in boldface type as "bursty", "stable", or "mixed" based on its incident edges. The dotted line shows a possible clustering of this graph based on these measures, with parts of the graph to the left and below the dotted line being bursty, and parts of the graph to the right and above the dotted line being stable. This illustrates one form of spatial correlation in this example.**

There are several particular questions that can be answered using these measures:

**Spatial Correlation.** It is possible that links behave similarly to other links that are close to them, by some measure of distance. Such behavior is known as *spatial correlation*. The dichotomy of burstiness or the statistical vectors proposed allow us to check for spatial correlation, defining distance from the networks perspective, such as a number of common friends distance measure. Figure 7 shows an example of partitioning a network based on spatial correlation. Clustering also follows naturally from spatial correlation.

**Information Flow over Time.** As links turn on and off, information may be transmitted between nodes at some times but not at others. Knowledge about when information can flow in the network can allow for a more detailed analysis of traditional information flow problems, such as information cascades.

# 6. MODEL FOR BURSTY LINK PATTERNS

We now consider a model that captures the qualitative behavior of individual links in the Research Institution e-mail network. Algorithm 1 provides an overview of the algorithm used to capture this model. Essentially, this algorithm takes a synthetically-generated bursty trace [10] $t$ and processes it, adding a random process at the time scale of size 2 units, based on the seed value from the overall trace. Thus, the output time series is twice as long as the input time series.

Figure 8 provides a preliminary evaluation of this model. We see that long-term trends are correctly captured, but that the real data has some higher variation in $y_j$, possibly due to non-stationarities that are inherent to human behavior. A potential extension to this model is introducing nonstationarities, such as multiplying the time series by sine wave, to increase the precision of the results.

**Data**: A bursty time series $t$ with $\alpha > 0$ for all or most time scales, including the largest scale.
**Result**: A time series $t'$ with the qualitative logscale characteristics of representative e-mail links.
$n := 2$;
**foreach** *Element in* $t$ **do**
  $r := 1 + \texttt{randn}()/n$; //`randn`() produces random number from standard normal distribution
  $filler := [r, r]$;
  $t'[n \cdot i + 1 : n \cdot (i + 1)] := filler. * trace(i + 1)$; // .* is elementwise multiplication
**end**
**return** $t'$;
**Algorithm 1**: Algorithm to model behavior on individual network links.

# 7. CONCLUSIONS AND FUTURE WORK

In this paper, we considered statistical techniques to model time-variations in social network links, possible implications and methods to apply these metrics to networks, and proposed a possible technique to model such time variations.

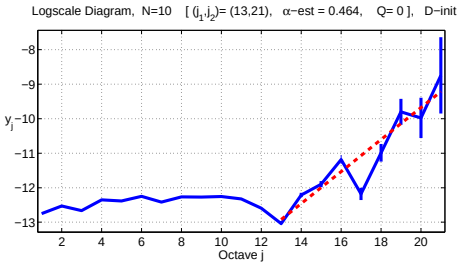
In future work, we hope to extend the various applications to network structure such as temporal variation and information flow, thus building on the ideas proposed here.

# 8. ACKNOWLEDGEMENTS

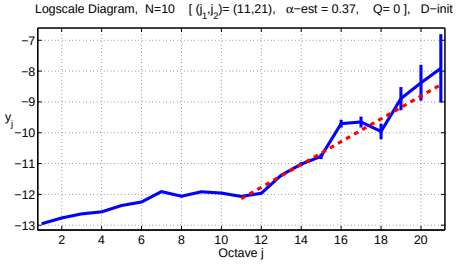
I appreciate extensive discussions with Jure Leskovec, Simla Ceyhan, and Borja Peleato.

# 9. REFERENCES

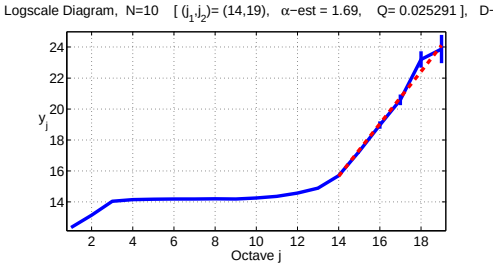
[1] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch. Wavelets for the analysis, estimation and synthesis of scaling data. *Self-Similar Network Traffic and Performance Evaluation*, pages 39–88, 2000.
[2] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.
[3] J. Eckmann, E. Moses, and D. Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences*, 101(40):14333, 2004.
[4] B. Klimt and Y. Yang. The enron corpus: A new dataset for email classification research. *Lecture notes in computer science*, 3201:217–226, 2004.
[5] G. Kossinets and D. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88, 2006.
[6] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking (ToN)*, 2(1):1–15, 1994.
[7] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506, 2009.
[8] R. Malmgren, D. Stouffer, A. Motter, and L. Amaral. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105(47):18153, 2008.
[9] W. Reed and M. Jorgensen. The double Pareto-lognormal distribution-a new parametric model for size distributions. *Communications in Statistics—Theory and Methods*, 33(8):1733–1754, 2004.
[10] T. Rusak and P. Levis. Burstiness and scaling in the structure of low-power wireless links. *ACM SIGMOBILE Mobile Computing and Communications Review*, 13(1):60–64, 2009.
[11] T. Rusak and P. Levis. Physically-based models of low-power wireless links using signal power simulation. *Elsevier Computer Networks*, 2009.
[12] S.-W. Seong, M. Nasielski, J. Seo, S. H. Debangsu Sengupta, S. K. Teh, R. Chu, B. Dodson, and M. S. Lam. The Architecture and Implementation of a Decentralized Social Networking Platform. *In submission*, October 2009.
[13] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskovec. Mobile call graphs: beyond power-law and lognormal distributions. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 596–604, 2008.
[14] A. Vázquez, J. Oliveira, Z. Dezsö, K. Goh, I. Kondor, and A. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):36127, 2006.




(a) Logscale diagram from a representative individual node (sender) in the Research Institution E-mail network.




(b) Logscale diagram from another representative individual node (sender) in the Research Institution E-mail network.




(c) Logscale diagram resulting from the proposed model.

**Figure 8: Evaluation for the model proposed, using logscale diagrams. The model can capture the qualitative biscaling behavior, i.e., a small positive slope followed by a steeper positively sloping section of the logscale diagram.**