REPORT DOCUMENTATION PAGE	Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)

2. REPORT TYPE
Technical Report

3. DATES COVERED (From - To)
-

4. TITLE AND SUBTITLE
A Graphical Exploration of the IkeNet E-mail Dataset

5a. CONTRACT NUMBER
W911NF-10-1-0472

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHORS
Andrea Bertozzi, Kate Coronges, Martin Short

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of California - Los Angeles
Regents of the University of California, Los Angeles
Office of Contract and Grant Administration
Los Angeles, CA 90095 -1406

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
58344-MA.15

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
This is a set of powerpoint slides regarding the IKENET database and a preliminary analysis.

15. SUBJECT TERMS
social network, internet traffic, self-exciting point process

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Andrea Bertozzi
UU	UU	UU	UU		19b. TELEPHONE NUMBER 310-825-4340

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **MAR 2012**	2. REPORT TYPE	3. DATES COVERED **00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE **A Graphical Exploration of the IkeNet E-mail Dataset**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California - Los Angeles,Los Angeles,CA,90095**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**15**	

Report Title

A Graphical Exploration of the IkeNet E-mail Dataset

ABSTRACT

This is a set of powerpoint slides regarding the IKENET database and a preliminary analysis.

A Graphical Exploration of the IkeNet E-mail Dataset

Martin Short, Andrea Bertozzi, and Kate Coronges (West Point)

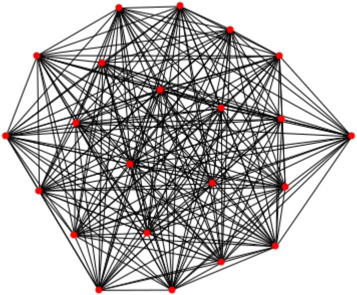
UCLA Applied Math

March 15, 2012

What is the IkeNet e-mail dataset?

- Certain cadets at West Point are given Blackberries in exchange for their willingness to have data about their communication activity logged and studied.
- We have a database on the e-mail communications within a network of 22 such students over ≈ 1 year, from May 2010 to May 2011.
- Note that only the e-mails sent *within* the network are included, not all e-mails sent by each subject.
- There are ≈ 8500 such emails, and each includes three pieces of information: sender, reciever, and timestamp.
- Today I will show you several plots made from this data, to hopefully elicit ideas about further avenues of exploration.

First, the network of e-mail traffic.



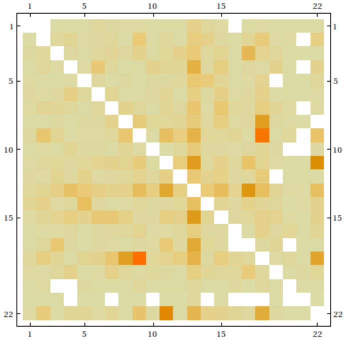


Figure 1: (Left) Dots represent the 22 subjects, and a line connects two dots if there is at least 1 correspondence between the two in our dataset. There is only 1 component, but it is not fully connected. (Right) A plot showing the number of e-mails sent from subject *i* (row) to subject *j* (column). Note this is a directed graph, and the matrix is not symmetric.

A histogram of e-mails per pair.

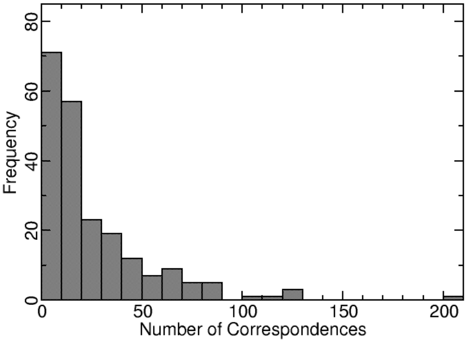
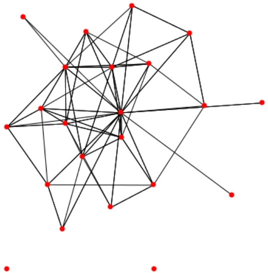


Figure 2: This is data from a symmetric version of the graph, and shows frequency of correspondences per pair. There are 5 pairs not shown here, each with many more correspondences. For example, pair $(9, 18)$ has 1032 messages!

Now, if we threshold the graph at 20 e-mails, we see some more detail.



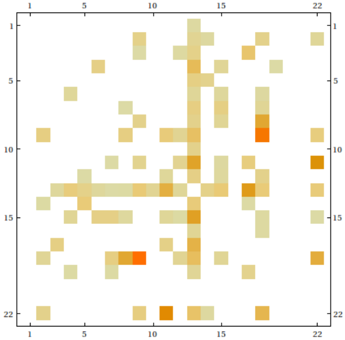


Figure 3: (Left) Here, we see that subject 13 begins to stand out as a central figure with by far the most "significant" connections. Subjects 20 and 21 are no longer a part of the network at all.

Now some temporal properties.

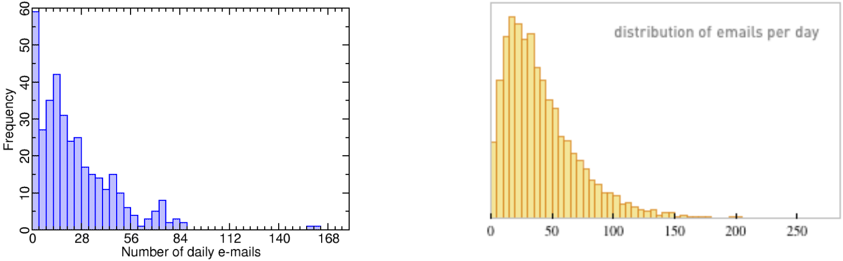


Figure 4: (Left) Histogram of e-mails sent per day in the IkeNet dataset. Note the large bar at < 4 e-mails – weekends? (Right) For fun, a similar plot from Stephen Wolfram, using his sent e-mails since 1989 (!). He is clearly more e-mail happy than the West Points cadets, but the general shape (omitting the origin) is similar. . .

But when are the e-mails sent?

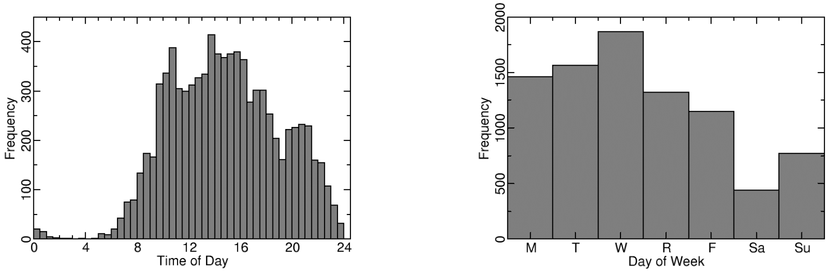


Figure 5: (Left) A histogram of when IkeNet e-mails were sent during the day. We clearly see a diurnal rhythm, modulated by lunch and dinner effects. (Right) Histogram of e-mails per weekday, clearly dropping off on the weekends.

We can see these cycles clearly in an auto-correlation analysis of the time series.

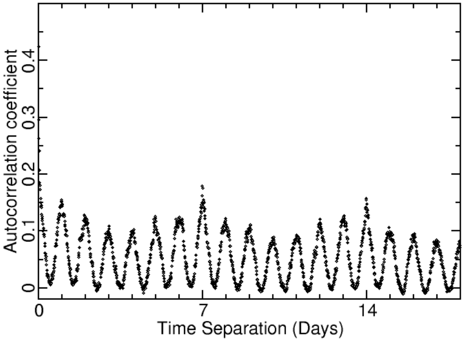


Figure 6: But, also note the large spike near the origin, indicating large correlation at very short timescales – self-excitation?

We can check for self-excitation using a “fixed-window” count.

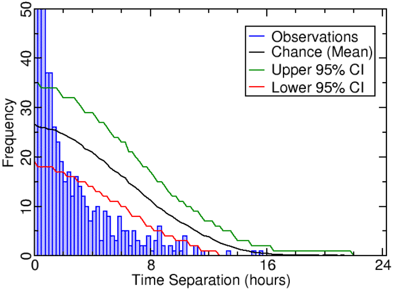


Figure 7: Here, we find all occurrences of subject pairs that exchanged exactly two e-mails on a calendar day (802 of these), then plot the frequency of time intervals between the two e-mails. The observations are vastly larger than chance at times less than around 1 hour. 311 are separated by less than 15 minutes.

We might try fitting a time series to a Hawkes process.

- Let's use the time series from pair $(9, 18)$, which is most prolific at 1032 messages.
- Fit the data to a process of the form

  $$\lambda(t) = \mu + k \sum_{t_i < t} \omega e^{-\omega(t - t_i)}$$

  using Maximum Likelihood Estimation.
- The best fit parameters are: $\mu = 0.054$ per hour, $k = 0.585$, and $\omega^{-1} = 0.099$ hours, with a log-likelihood of -1303.6.
- These parameters tell us that there were around 428 background events for this pair, and 604 excited events. That's a lot of excitation...

We can also fit non-parametrically using EM (as in Mohler *et al.*, 2011 *JASA*)

Here $\lambda(t) = \mu(t) + \sum_{t_i < t} g(t - t_i)$.

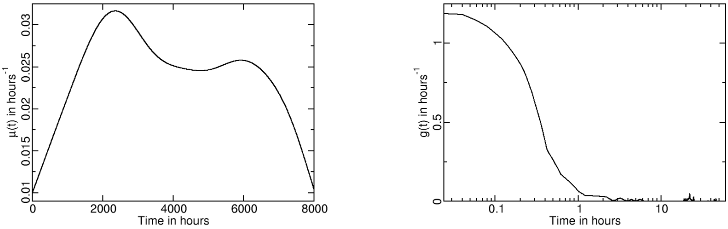


Figure 8: Here we show KDEs of the background $\mu(t)$ and excited kernel $g(t)$ for pair $(9, 18)$. These kernels give roughly 202 background events, and 1,100 excited events. Log-likelihood is -1379.4, though, which is worse than Hawkes.

What might we explore next?

- Do a more careful EM analysis (MPLE?).
- Build daily and weekly rhythms directly into the EM or Hawkes process, since a lot of information is there.
- Explore data from other pairs (i, j), and perhaps include multi-party interactions.
- Look (much) more deeply into the graph structure. Perhaps using some of Uminsky's coalition finding techniques?
- Try to obtain more data from different sources (GMail?) on frequency of emails sent per day, to explore perhaps a simple model that explains similarities (and differences) between IkeNet and Wolfram.
- All this, and much, much more...