**Title:** A Graphical Exploration of the IkeNet E-mail Dataset

**Authors:** Andrea Bertozzi, Kate Coronges, Martin Short

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

**Subject Terms:**
Social network, internet traffic, self-exciting point process

**Security Classification:**
Unclassified

**Distribution Availability Statement:**
Approved for public release; distribution is unlimited.
A Graphical Exploration of the IkeNet E-mail Dataset

Title: A Graphical Exploration of the IkeNet E-mail Dataset

Author: Unknown

Performing Organization: University of California - Los Angeles, Los Angeles, CA, 90095

Report Date: MAR 2012

Report Type: 00-00-2012 to 00-00-2012

Distribution/Availability Statement: Approved for public release; distribution unlimited

Subject Terms: Unknown

Security Classification of: Unclassified

Limitation of Abstract: Same as Report (SAR)

Number of Pages: 15
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.
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 \( \approx 1 \) 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 \( \approx 8500 \) such emails, and each includes three pieces of information: sender, receiver, 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...