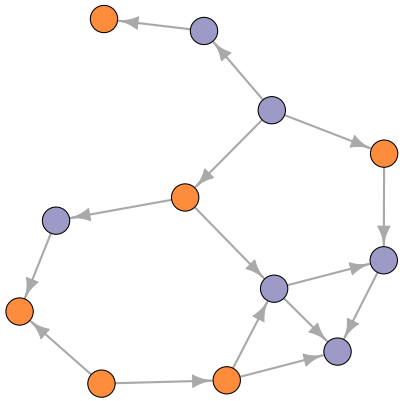
Point Process Network Models

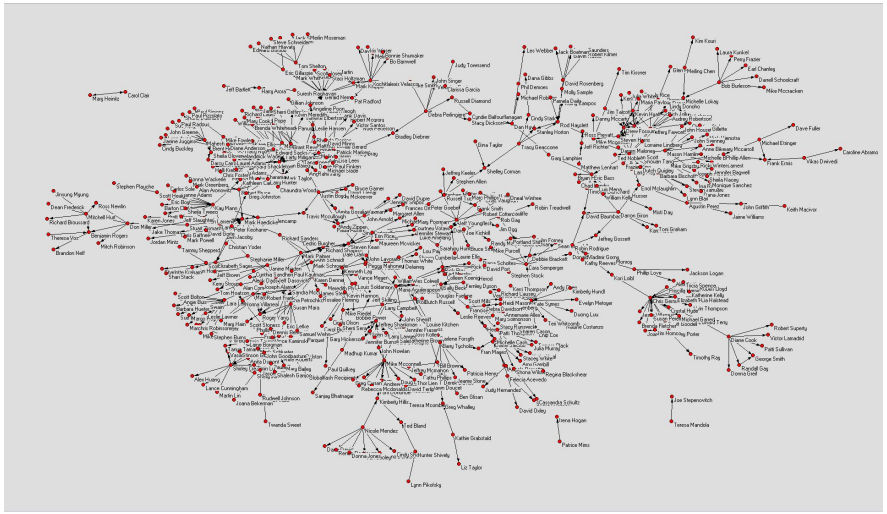
Patrick O. Perry & Patrick J. Wolfe
Harvard University

Graph Exploitation Symposium
MIT Lincoln Laboratory, 9th August 2011

Prelude

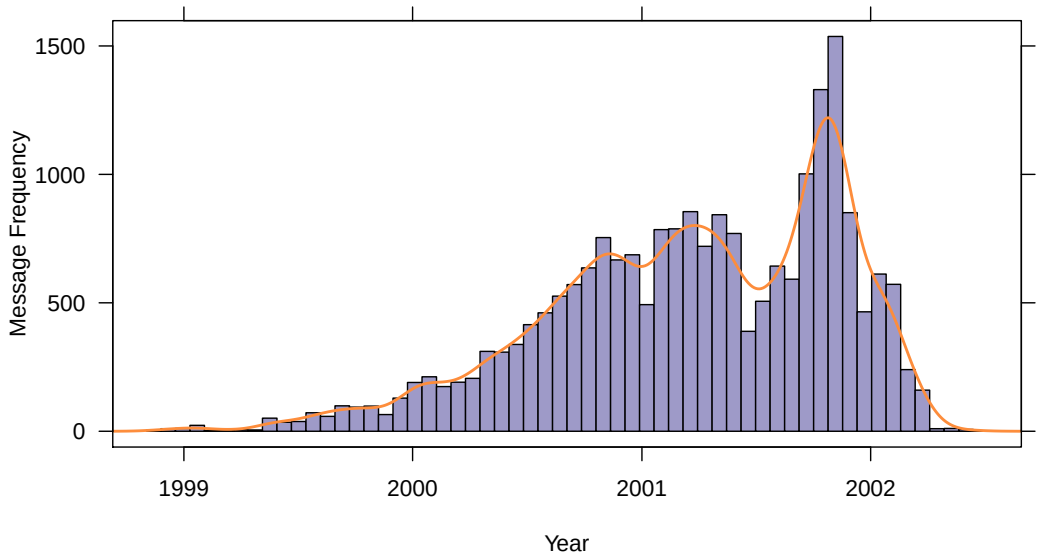
Motivation





- Networks are a natural way to represent high-dimensional yet sparse correlation structure
- Data often take the form of repeated pairwise interactions
- A multivariate point process representation is simple, flexible, and useful in this setting
- We motivate this approach through the analysis of a corporate e-mail data set

Point Process Approach



- Interaction data are often summarized as counts
- 'New social media', online messaging, etc...
- Network comprises sets $\mathcal{I}$ of 'sender' nodes and $\mathcal{J}$ of receivers, observed on a time interval $[0, T]$
- Interactions (e.g., email exchanges) may have single or multiple receivers

Simplest Version

- Suppose we assume constant-rate Poisson 'send' processes, & constant-rate selection of a single receiver for each message
- This reduces to fitting $2N$ node-specific parameters, for a directed graph on N nodes
- ML estimates are obtainable in closed form (self-loops) or iteratively; Fisher information also available
- Doesn't fit real-world data very well at all...(but gives rise to residuals-based analysis, a.k.a. 'network modularity')

Introduction

A Corporate Email Network

The Enron corpus: a large collection of email messages sent within the company between November 1998 and June 2002

21,635 messages
156 employees

A Typical Email Message

```
Message-ID:
<7303996.1075860726914.JavaMail.evans@thyme>
Date: Wed, 10 Oct 2001 08:51:16 -0700 (PDT)
From: kenneth.lay@enron.com
To: benjamin.rogers@enron.com
Subject: RE: Power Trading Group
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit

Ben -

I likewise was glad to see you.  Sorry we didn't
have a chance to talk.

Good to hear you're doing well.  You're with a great
group and, yes, the company will soon be doing a lot
better.

Thanks,

Ken
```

Homophily in the Network?

- **Question:** Is group membership predictive of interaction?
  - Gender, Department, Seniority
- **Answer(?):** Contingency table analysis, homogeneity assumptions are violated:
  - Dependence, Time variation, Multi-way interactions

Other questions: Are past interactions predictive of future ones? Does this effect vary over time? How should multiple-receiver interactions be handled? Can these be treated as multiple pairwise interactions? . . .

Contingency Table Analysis

	Legal Jr	Legal Sr	Trading Jr	Trading Sr	Other Jr	Other Sr
Legal Jr	-0.07	2.8	-1.91	2.88	-0.3	-0.4
Legal Sr	1.39	0.3	2.58	-0.15	-1.0	0.9
Trading Jr	-0.15	-Inf	-0.76	1.05	1.3	2.3
Trading Sr	4.41	0.6	0.28	-0.07	0.7	0.1
Other Jr	0.36	0.9	-0.01	1.44	1.0	-1.3
Other Sr	0.82	1.6	1.23	-0.30	-0.4	0.6

	Legal Jr	Legal Sr	Trading Jr	Trading Sr	Other Jr	Other Sr
Legal Jr	""	"++++"	"----"	"++++"	""	""
Legal Sr	"++++"	"+++"	"++"	""	""	"++++"
Trading Jr	""	"---"	"---"	"+++"	"++"	""
Trading Sr	"++++"	""	""	""	"+++"	""
Other Jr	""	""	""	"++++"	"++++"	"----"
Other Sr	"+"	"++++"	"++"	""	"-"	"++++"

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Positive log-odds indicates homophily ('birds of a feather')
- Fisher's exact test yields significance levels
- Validity?

Dependence

Date: Wed, November 7, 2001 8:34 AM
From: Webb, Jay
To: Kitchen, Louise
Subject: Fw: 8:30 am trade count
Hi Louise,

We are having a typical trading pace so far today. It is too early to tell if any counterparty is really cutting back. Like yesterday, however, Aquila is buying longer dated physical gas and selling spot gas...

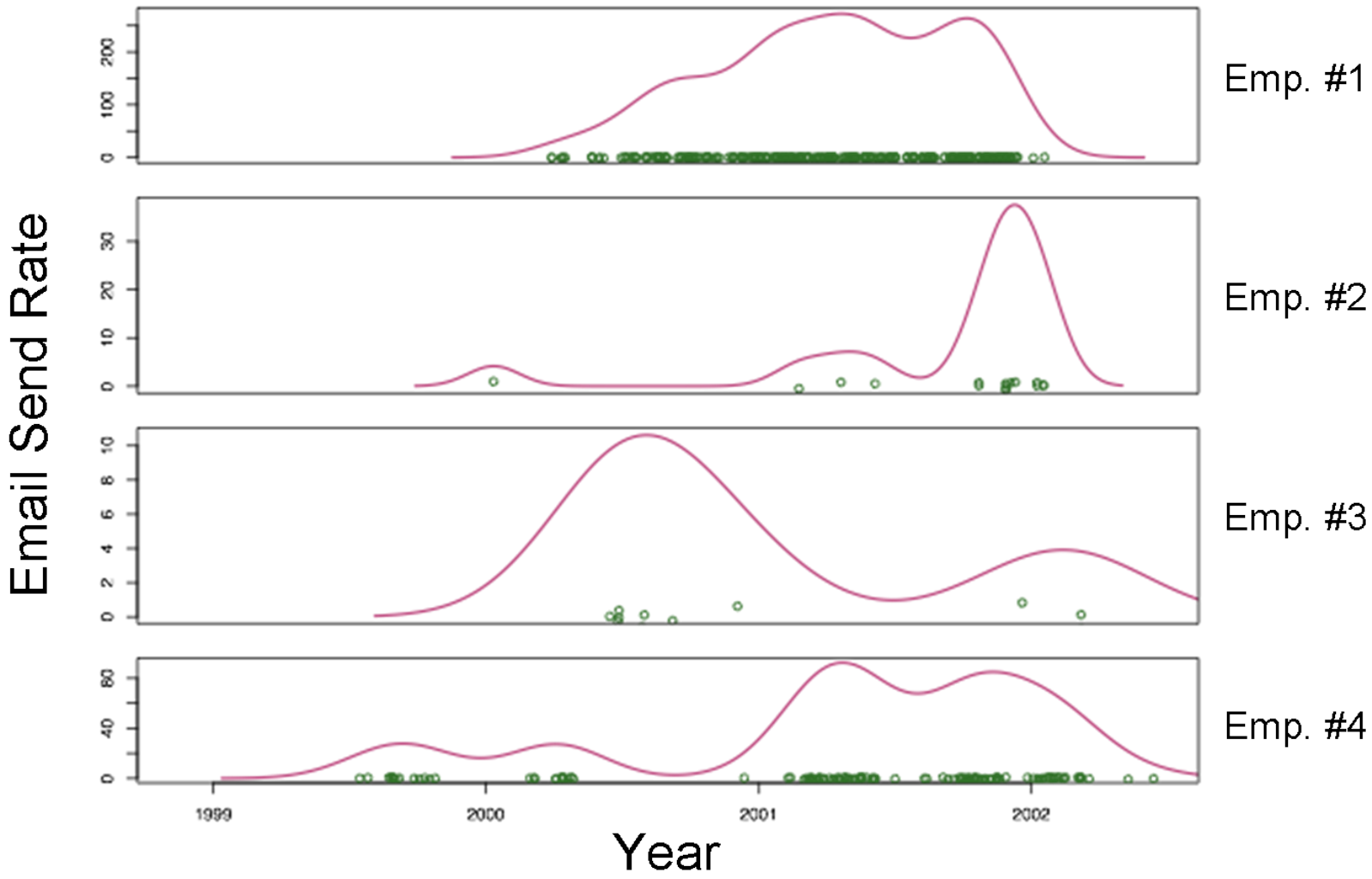
Date: Wed, November 7, 2001 10:14 AM
From: Kitchen, Louise
To: Arnold, John; Shilvey, Hunder; Neil, Scott; Martin, Tom; Grigsby, Mike
Subject: Fw: 8:30 am trade count

Note aquila.

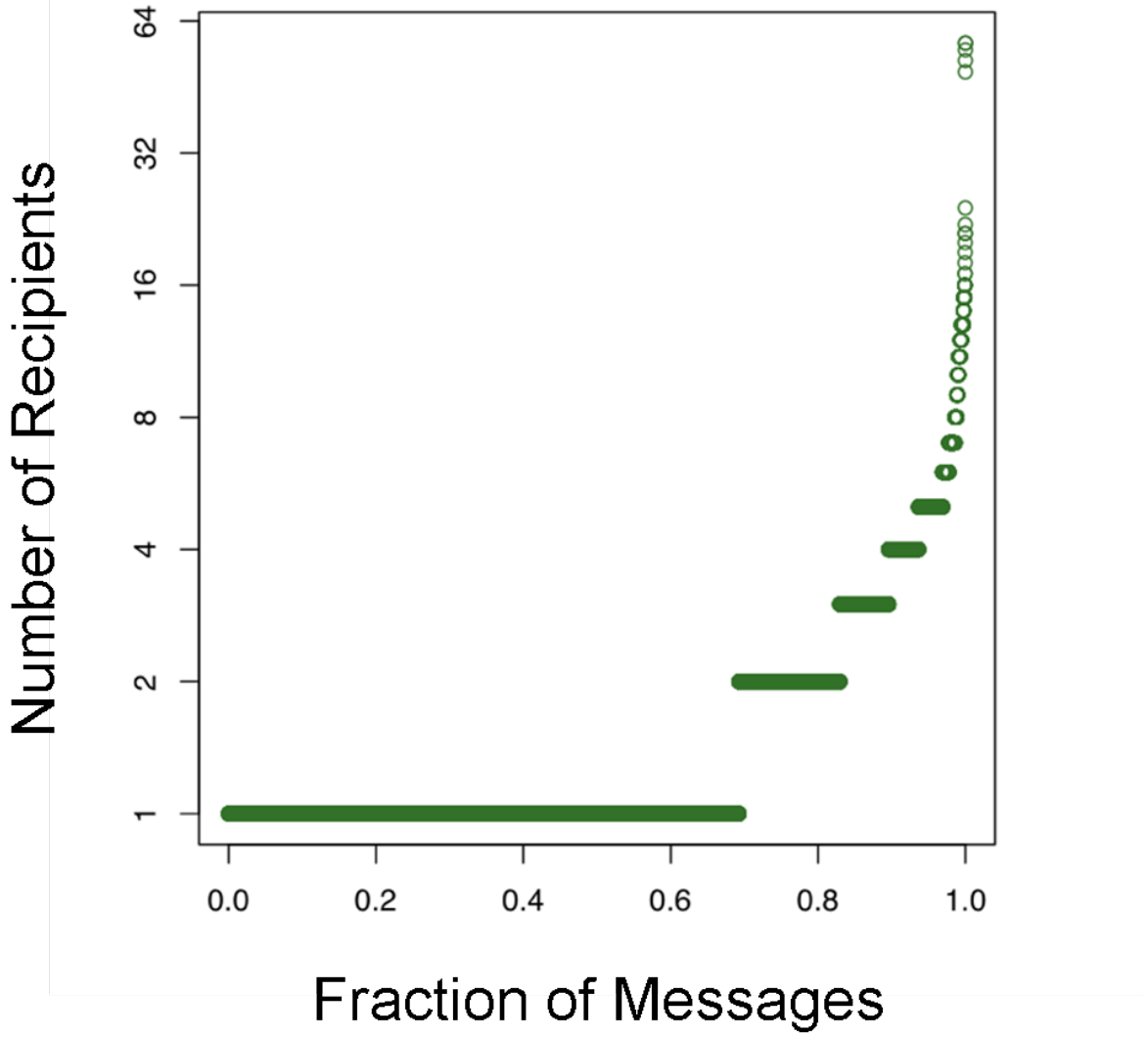
Date: Wed, November 7 2001 8:19 AM
From: Arnold, John
To: Kitchen, Loise; Webb, Jay
Subject: RE: 8:30 am trade count

fyi : Having more and more counterparties that will only deal on one side of my market.

Varying Rates



Multiple Recipients



Modeling

Proportional Intensity Model

Model pairwise interactions $i \to j$ via stochastic intensity $\lambda_t(i,j)$:

$$\lambda_t(i,j)\,dt = \Pr\{\text{interaction } i \to j \text{ occurs in time } [t, t+dt)\}.$$

Sender i interacts with receiver j at a baseline rate $\bar{\lambda}_t(i)$ modulated up or down according to the pair's covariate vector, $x_t(i,j)$:

$$\lambda_t(i,j) = \bar{\lambda}_t(i) \cdot \exp\{\beta_0^T x_t(i,j)\} \cdot 1\{j \in \mathcal{J}_t(i)\}.$$

- $\mathcal{J}_t(i)$ is the receiver set of sender i at time t
- $\bar{\lambda}_t(i)$ denotes the baseline intensity of sender i
- $x_t(i,j) \in \mathbb{R}^p$ comprises covariates; coefficient vector β_0

Covariate Possibilities

- Static Covariates: same gender, same dept, same seniority

$$1\{i \text{ and } j \text{ belong to the same group}\}$$

- Dynamic Covariates: received from j last minute, hour, day, week, month, etc.

$$1\{\text{interaction } j \longrightarrow i \text{ occurred in } [t - \delta_l, t)\}$$

Any process depending only on the past is a valid covariate; e.g.,

$$1\{\text{for some } k, \text{ interactions } i \longrightarrow k \text{ and } k \longrightarrow j \text{ occurred in } [t - \delta_l, t)\}$$

Inference

Treat $\bar{\lambda}_t(i)$ as a nuisance parameter (Cox's partial likelihood):

a) Log partial likelihood at time t, evaluated at β:

$$\log PL_t(\beta) = \sum_{t_m \le t} \left\{ \beta^T x_{t_m}(i_m, j_m) - \log \Big[\sum_{j \in \mathcal{J}_{t_m}(i_m)} \exp\{\beta^T x_{t_m}(i_m, j)\} \Big] \right\}$$

b) Approximate "multicast" likelihood:

$$\log \widetilde{PL}_t(\beta) = \sum_{t_m \le t} \left\{ \sum_{j \in J_m} \beta^T x_{t_m}(i_m, j) - |J_m| \log \Big[\sum_{j \in \mathcal{J}_{t_m}(i_m)} \exp\{\beta^T x_{t_m}(i_m, j)\} \Big] \right\}$$

NB: Maximizing $\log \widetilde{PL}_t(\cdot)$ instead of $\log PL_t(\cdot)$ introduces bias

Asymptotics

Different asymptotic regime than traditional proportional hazards

For pairwise interactions, under suitable regularity conditions:

Theorem (Perry & W, 2010)

As the number n of interactions grows,

i) The maximum likelihood estimator $\hat{\beta}_n$ of β_0 is consistent; i.e., it converges in probability to β_0;

ii) The quantity $\sqrt{n}\,(\hat{\beta}_n - \beta_0)$ converges in distribution to a zero-mean Normal random variable whose covariance can also be consistently estimated.

Results also extend to the case of multiple recipients (more work)

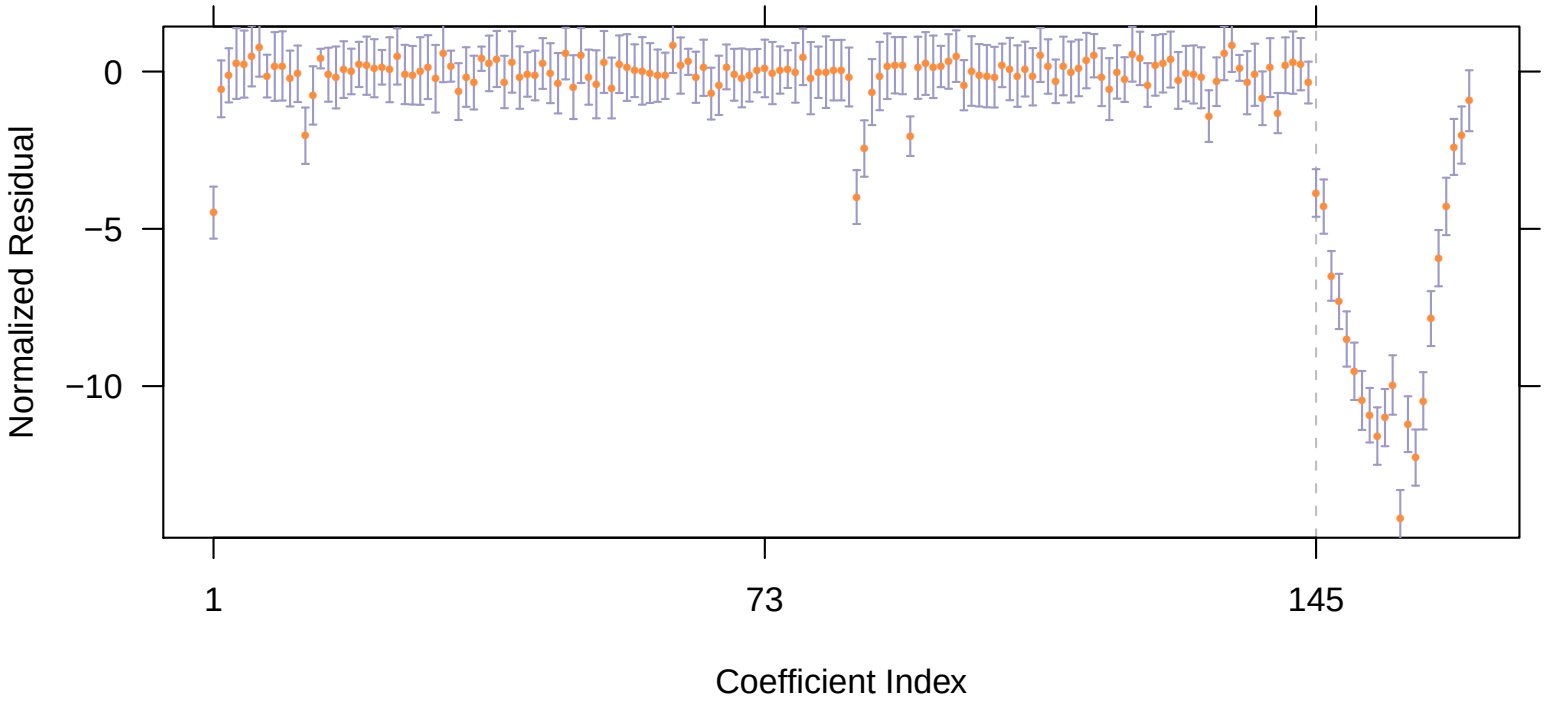
Results

Goodness of Fit

Term	Df	Deviance	Resid. Df	Resid. Dev
Null			35567	358759
Static	132	63809	35435	294950
Dynamic	21	86831	35414	208119

- Group-level (static) effects account for 18% of the residual deviance and reciprocation (dynamic) effects account for 24%
- Residual deviance is about 6$\times$ the residual degrees of freedom (overdispersion)

Multicast Bias Correction



Bootstrap residuals normalized by standard errors

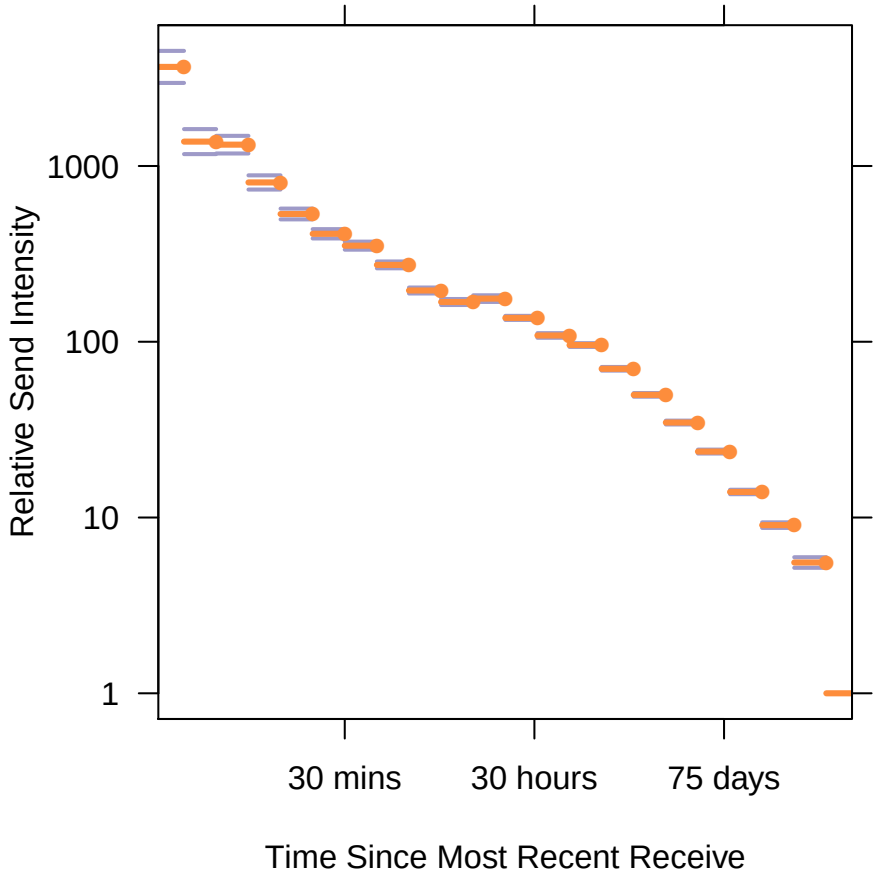
- Note (correctable) negative bias in the coefficient estimates

Static Effects

	Sender					
Receiver	FLJ	FLS	FTJ	FTS	FOJ	FOS
FLJ	**3.31** (0.11)	3.06 (0.25)	1.55 (0.30)	1.25 (0.35)	0.29 (0.08)	0.69 (0.10)
FLS	2.55 (0.11)	4.21 (0.41)	0.15 (0.14)	0.76 (0.37)	0.39 (0.15)	1.00 (0.20)
FTJ	0.43 (0.06)	0.61 (0.24)	1.36 (0.33)	1.39 (0.38)	0.46 (0.12)	1.00 (0.29)
FTS	0.81 (0.12)	0.49 (0.14)	**4.34** (1.11)	**2.53** (0.65)	2.22 (0.28)	0.19 (0.10)
FOJ	0.47 (0.05)	0.14 (0.06)	0.96 (0.29)	1.54 (0.27)	**2.86** (0.23)	1.92 (0.18)
FOS	0.99 (0.07)	1.11 (0.19)	0.11 (0.09)	0.70 (0.34)	1.46 (0.16)	**3.84** (0.30)
MLJ	1.41 (0.10)	1.03 (0.28)	3.62 (0.69)	0.65 (0.42)	1.48 (0.32)	0.76 (0.15)
MLS	3.07 (0.11)	**4.48** (0.35)	1.15 (0.45)	0.65 (0.21)	0.35 (0.09)	1.48 (0.14)
MTJ	0.70 (0.06)	1.42 (0.18)	2.10 (0.38)	0.61 (0.17)	0.66 (0.10)	0.33 (0.08)
MTS	0.61 (0.05)	1.32 (0.15)	2.68 (0.41)	2.16 (0.29)	1.58 (0.14)	0.99 (0.10)
MOJ	0.47 (0.04)	0.27 (0.05)	2.16 (0.35)	1.34 (0.21)	1.62 (0.13)	0.75 (0.08)
MOS	0.86 (0.06)	0.71 (0.10)	0.13 (0.10)	0.37 (0.14)	2.39 (0.20)	3.74 (0.28)
	MLJ	MLS	MTJ	MTS	MOJ	MOS
FLJ	2.21 (0.24)	2.97 (0.24)	0.27 (0.07)	0.08 (0.02)	0.14 (0.06)	0.70 (0.12)
FLS	0.45 (0.24)	2.46 (0.21)	2.33 (0.35)	0.42 (0.08)	0.11 (0.07)	0.38 (0.09)
FTJ	2.14 (0.32)	0.06 (0.05)	**6.19** (0.56)	**2.26** (0.17)	3.13 (0.35)	0.06 (0.05)
FTS	0.44 (0.21)	0.48 (0.10)	1.18 (0.24)	2.00 (0.17)	0.51 (0.10)	0.39 (0.15)
FOJ	0.39 (0.10)	0.26 (0.06)	0.27 (0.07)	1.01 (0.09)	2.07 (0.19)	3.32 (0.35)
FOS	1.80 (0.30)	2.35 (0.21)	0.13 (0.06)	1.68 (0.14)	2.13 (0.26)	2.24 (0.22)
MLJ	0.94 (0.26)	1.52 (0.22)	0.70 (0.20)	2.24 (0.20)	1.04 (0.28)	2.26 (0.41)
MLS	1.79 (0.22)	**6.69** (0.51)	0.51 (0.12)	0.76 (0.07)	0.31 (0.09)	2.13 (0.21)
MTJ	0.67 (0.14)	1.33 (0.14)	2.33 (0.21)	1.00 (0.07)	2.64 (0.25)	0.58 (0.10)
MTS	**2.80** (0.28)	0.63 (0.06)	2.93 (0.23)	2.18 (0.10)	2.61 (0.24)	2.26 (0.22)
MOJ	0.69 (0.14)	0.15 (0.03)	3.60 (0.28)	1.19 (0.07)	**4.26** (0.36)	0.92 (0.11)
MOS	0.69 (0.12)	5.75 (0.45)	0.71 (0.10)	0.84 (0.06)	0.96 (0.12)	**3.53** (0.33)

Static effects as a function of shared sender/receiver groups

Reciprocation Effects



Estimated 'reciprocation' effects, as multiple of baseline rate

Conclusion

Big Picture

- Many network data sets take the form of repeated interactions
- Point process representation is simple, flexible, and useful
- Modeling message exchanges in a corporate e-mail network
  - Enables evaluation of which characteristics & behaviors appear predictive of interaction
  - Enables quantitative description of dynamic effects (e.g., reciprocation)

NSF-DMS/MSBS/CISE, DARPA, ONR, ARO MURI and PECASE support gratefully acknowledged.

Thanks also to the organizers for many fruitful technical discussions.