# Scalable Methods for the Analysis of Network-Based Data

MURI Project: University of California, Irvine

Project Meeting

May 25th 2010

Principal Investigator: Padhraic Smyth

# Today's Meeting

- Goals
  - Review our research progress
  - Discussion, questions, interaction
  - Feedback from visitors

- Format
  - Introduction
  - Research talks
    - 20 and 30 minute slots
    - 5 mins at end for questions/discussion
  - Question/discussion encouraged during talks
  - Several breaks for discussion

# Project TimeLine

- Project start/end
  - Start date: May 1 2008
  - End date: April 30 2011/2013
- Meetings
  - Nov 2008: All-Hands Kickoff Meeting
  - April 2009: Working Meeting
  - August 2009: Working Meeting
  - December 2009: All-Hands Annual Review
  - May 2010: Working Meeting

# MURI Investigators

Padhraic Smyth UCI

David Eppstein UCI

Carter Butts UCI

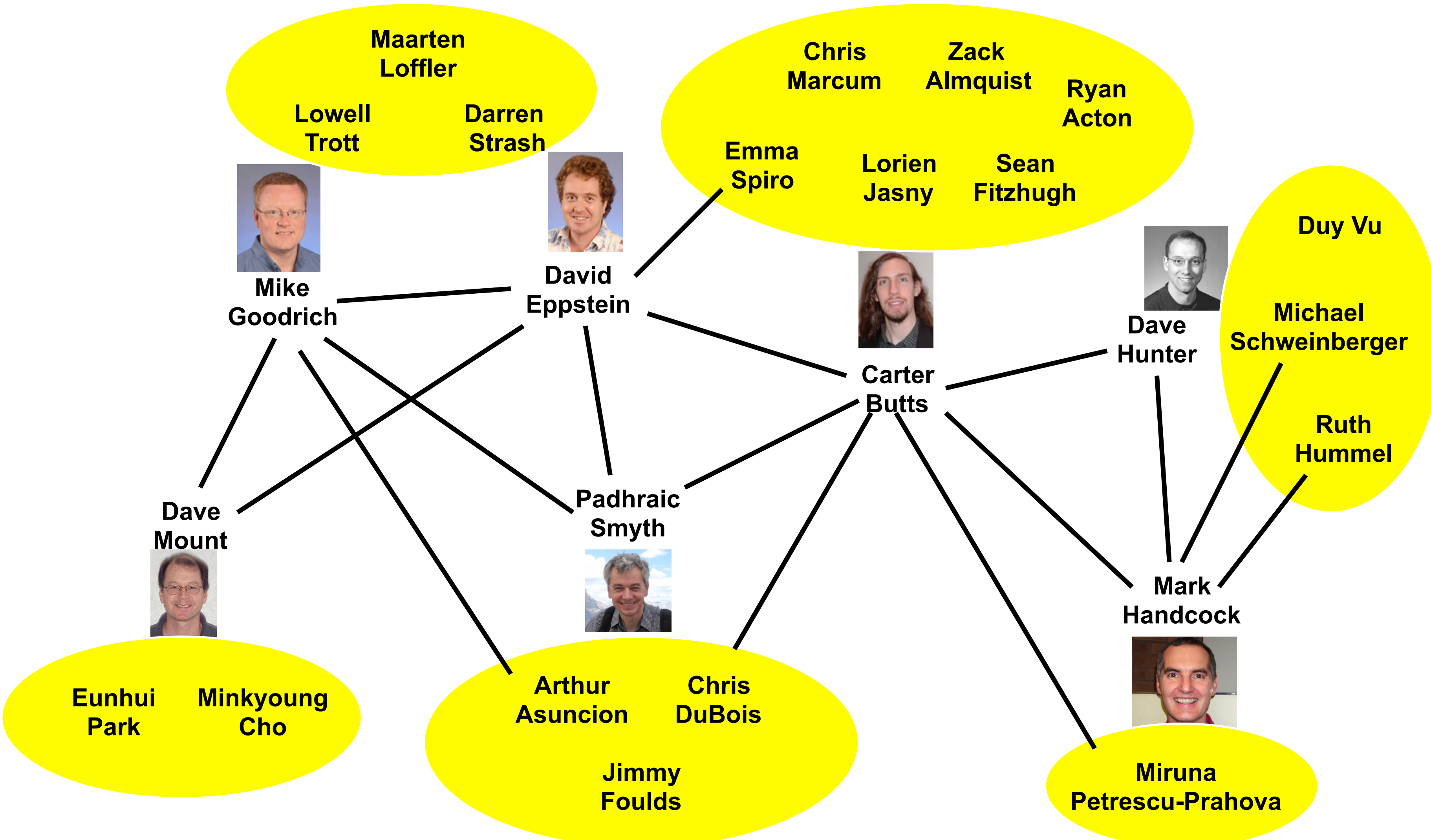
Michael Goodrich UCI

Mark Handcock
UCLA

Dave Mount
U Maryland

Dave Hunter
Penn State

# Collaboration Network



# Graduate Student Progress

Highlights

- Presenting talks at multiple international conferences this summer
    - Sunbelt International Social Networks conference (Jasny, Spiro, Fitzhugh, Almquist)
    - ACM SIGKDD Conference (DuBois)
    - American Sociological Meeting (Marcum, Jasny, Spiro, Fitzhugh, Almquist)
    - + more
- Workshop organization/instruction
    - Political Networks Conference (Spiro, Fitzhugh, Almquist)
- Summer school on social network analysis
    - DuBois and Almquist received scholarships to attend
- Faculty position at U Mass Amherst (Acton)
- Best paper awards or nominations (Spiro, Hummel)
- National fellowships (DuBois, Asuncion)

# Publications

**Fundamentals of Exponential Random Graph Models and Network Analysis**

Revisiting the foundations if network analysis, C.T. Butts, *Science*, 325, 414-416, 2009.

**Scalable Algorithms for Statistical Network Modeling**

The h-index of a graph and its application to dynamic subgraph statistics, D. Eppstein & E. S. Spiro. *Proceedings of Algorithms and Data Structures Symposium (WADS),* Springer-Verlag, Lecture Notes in Computer Science 5664, pp. 278-289, 2009.

A stepwise algorithm for fitting ERGMS, R.M. Hummel, M.S. Handcock, D.R. Hunter, Penn State Department of Statistics Technical Report 10-03, 2010.

Learning with blocks: composite likelihood and contrastive divergence, A. Asuncion, Q. Liu, A. T. Ihler, and P. Smyth. *Proceedings of the 13th International Conference on AI and Statistics*, May 2010.

Particle-filtered MCMC-MLE with connections to contrastive divergence, A. Asuncion, Q. Liu, A. Ihler, and P. Smyth, *Proceedings of the 27th International Conference on Machine Learning (ICML),* to appear, 2010.

# Publications

**Geometric and Spatial Embedding Methods**

Space-time tradeoffs for approximate nearest-neighbor searching, S. Arya, T. Malamatos, and D. M. Mount, *Journal of the ACM*, 57 (2009), 1-54.

Approximate range-searching: the absolute model, G. D. da Fonseca and D. M. Mount, *Computational Geometry*, 43:4, 434—444, 2010.

Approximation algorithm for the kinetic robust center algorithm, S. Friedler and D. M. Mount, *Computational Geometry*, 43(6-7), 572-586, 2010.

Maintaining nets and net trees under incremental motion, M. Cho, D. M. Mount, and E. Park, in *Proceedings of the 20th Intl. Symp. on Algorithms and Computation (ISAAC 2009),* 1134-1143, 2009

Particle-based variational inference for continuous systems, A. T. Ihler, A. J. Frank, P. Smyth, *Proceedings of the 22nd Neural Information Processing Conference (NIPS)*, Dec 2009.

Succient greedy geometric routing in the Euclidean plane, M. T. Goodrich and D. Strash, in *Proceedings of the 20th Intl. Symp. on Algorithms and Computation* (ISAAC 2009), 2009 (to appear).

The effect of corners on  the complexity of approximate range searching, S. Arya, T. Malamatos, and D. M. Mount, *Discrete and Computational Geometry*, 41 (2009), 398-443.

Compressing kinetic data from sensor networks, S. Friedler and D. M. Mount, *Algorithmic Aspects of Wireless Sensor Networks (ALGOSENSORS 2009)*, Springer Lecture Notes LNCS 5804, 2009, 191-202.

# Publications

**Dynamic and Relational Event Models**

Change and external events in computer-mediated citation networks: English language Weblogs and the 2004 electoral cycle, C. T. Butts and B. R. Cross, *Journal of Social Structure*, 10, 2010.

Modeling relational events via latent classes, C. DuBois and P. Smyth, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 2010, in press.

**Statistical Modeling of Text and Networks**

Distributed algorithms for topic models, D. Newman, A. Asuncion, P. Smyth, M. Welling, *Journal of Machine Learning Research*, 1801-1828, 2009.

Asynchronous distributed estimation of topic models for document analysis, , A. Asuncion, P. Smyth, M. Welling, *Statistical Methodology*, in press, 2010.

**Measurement of Large Scale Networks**

A walk in Facebook: uniform sampling of users in online social networks, M. Gjoka, M. Kurant, C. T. Butts, A. Markopoulou, *Proceedings of the IEEE Infocom Conference*, 2010

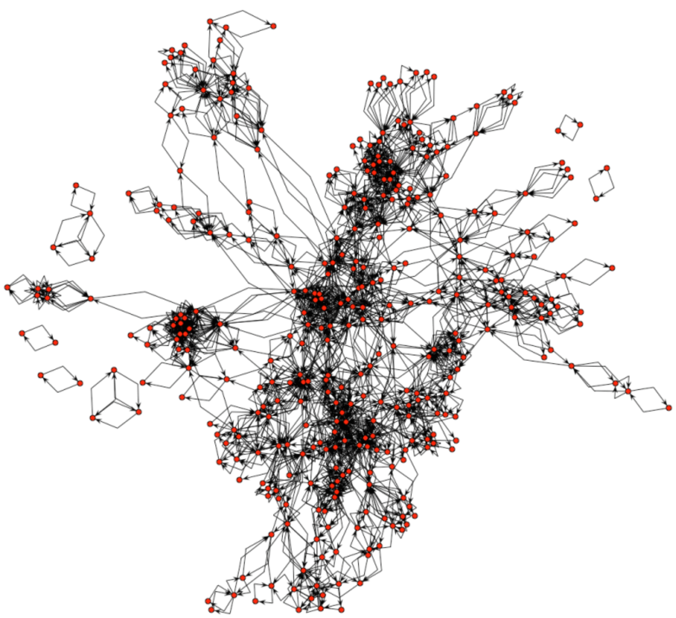
Data

Statistical Models

Scalable Algorithms

Evaluation

Software and Applications



# Statistical Modeling of Network Data

Statistics = principled approach for inference from noisy data

Integration of different sources of information

- e.g., combining edge information with node attributes

Basis for optimal prediction

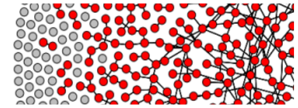
- computation of conditional probabilities/expectation

Principles for handling noisy measurements

- e.g., noisy and missing edges

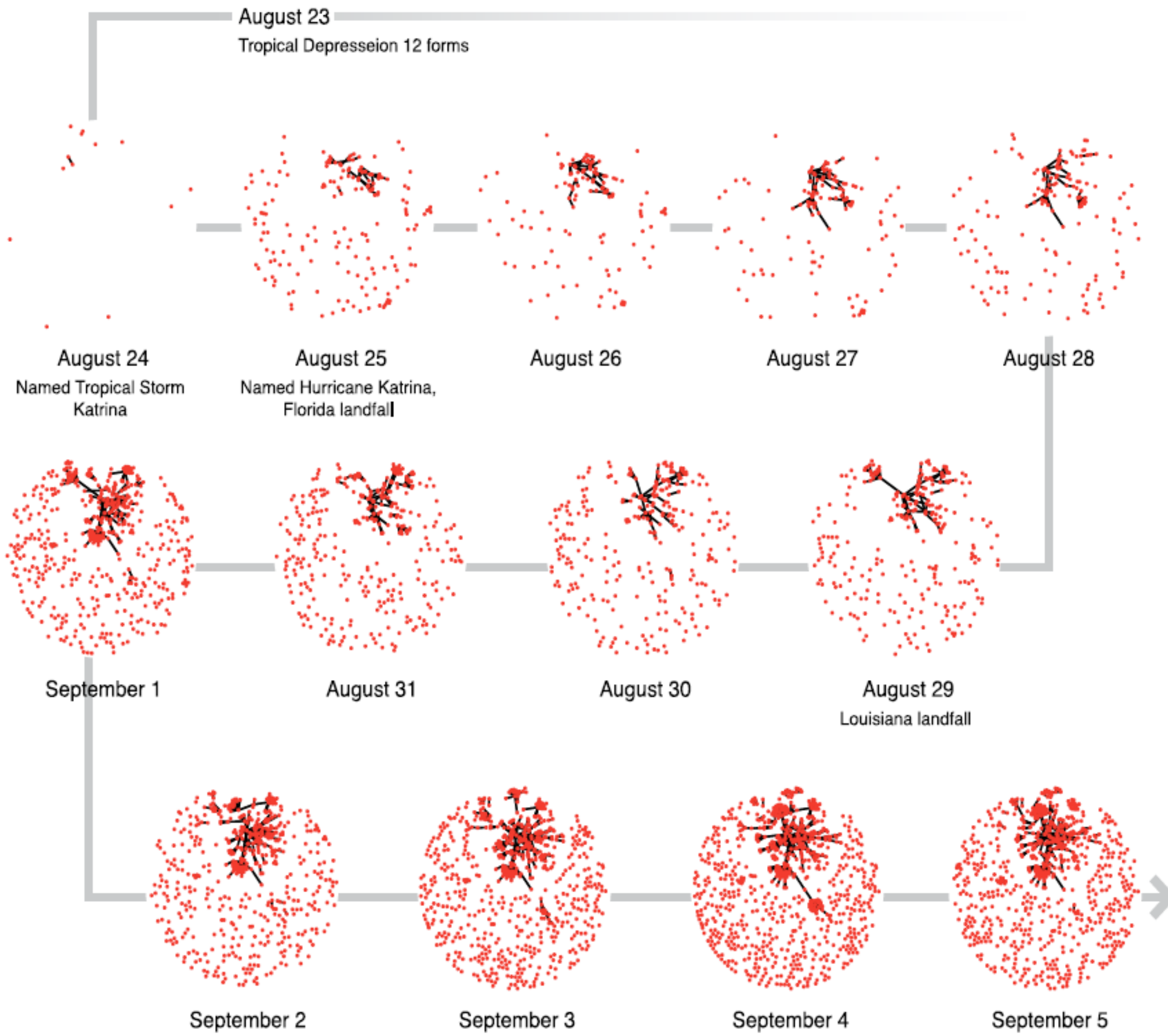
Quantification of uncertainty

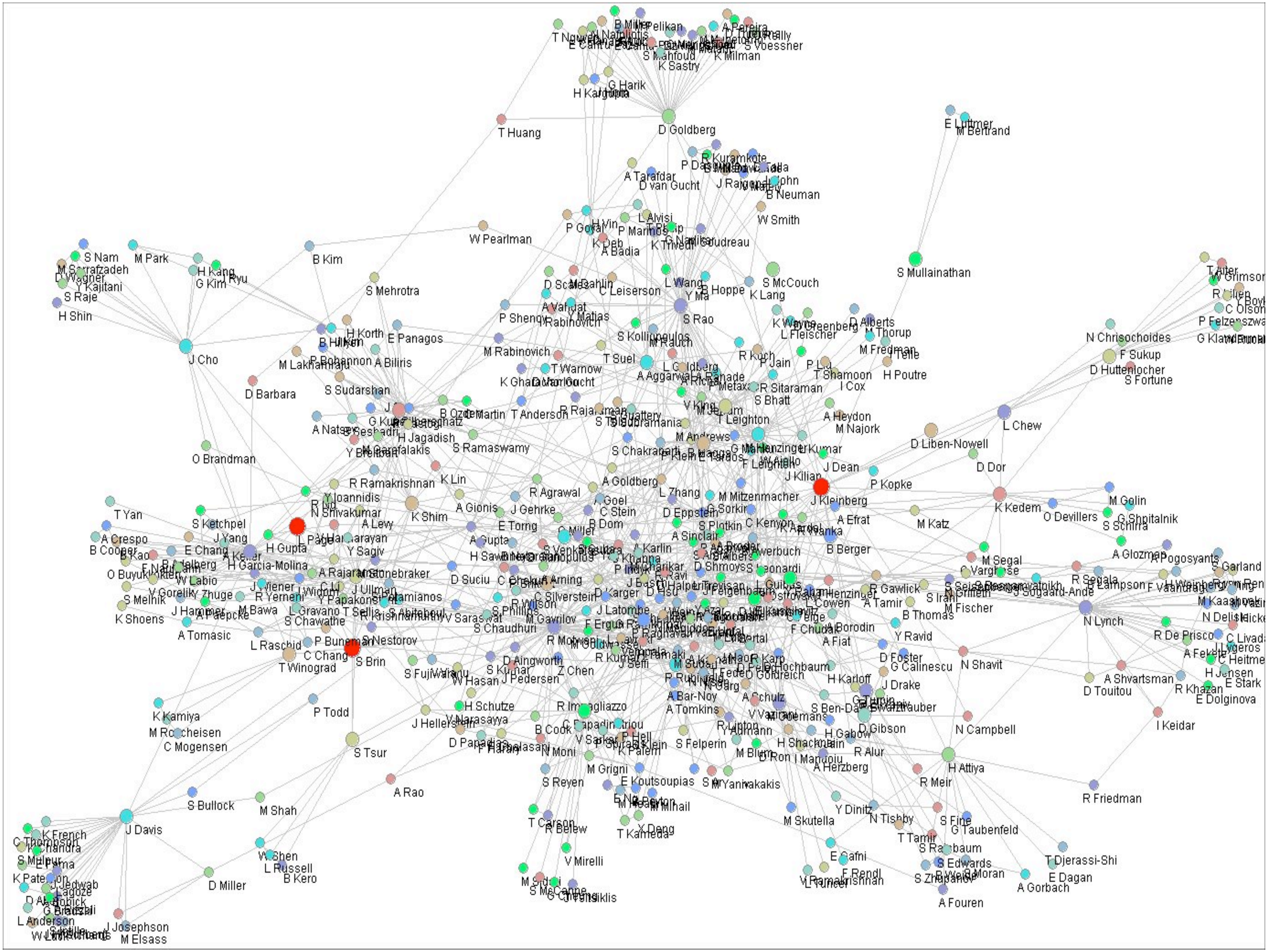
- e.g., how likely is it that network behavior has changed?



# Limitations of Prior Work

- Network data over time
  - Relatively little work on dynamic network data
- Heterogeneous data
  - e.g., few techniques for incorporating text, spatial information, etc, into network models
- Computational tractability
  - Many network modeling algorithms scale exponentially in the number of nodes *n*
  - Limits practical network sizes to order of $n = 100$ nodes


August 23
Tropical Depresseion 12 forms
August 24
Named Tropical Storm Katrina
August 25
Named Hurricane Katrina, Florida landfall
August 26
August 27
August 28
August 29
Louisiana landfall
August 30
August 31
September 1
September 2
September 3
September 4
September 5



T Nguyen
S Mahfoud
K Sastry
K Milman
S Voessner
A Pereira
G Harik
D Goldberg
T Huang
E Luttmer
M Bertrand
A Tarafdar
D van Gucht
B Neuman
W Smith
W Pearlman
L Alvisi
S Nam
M Park
M Sarrafzadeh
D Wagner
Y Kajitani
S Raje
H Shin
H Kang
G Kim
J Ryu
B Kim
S Mehrotra
S Mullainathan
S McCouch
L Wang
Y Ma
B Hoppe
K Lang
C Leiserson
S Rao
P Shenoy
J Cho
H Korth
E Panagos
A Biliris
M Lakhamraju
D Barbara
S Sudarshan
M Rabinovich
T Warnow
T Suel
S Kollipoulos
M Rauch
K Wang
L Fleischer
D Alberts
M Thorup
M Fredman
H Poutre
I Cox
T Shamoon
T Leighton
N Chrisochoides
F Sukup
D Huttenlocher
S Fortune
P Felzenszwalb
C Olson
L Chew
D Liben-Nowell
D Dor
J Dean
A Heydon
M Najork
R Sitaraman
S Bhatt
T Anderson
A Natsev
H Jagadish
S Ramaswamy
O Brandman
R Ramakrishnan
K Lin
Y Ioannidis
N Shivakumar
A Lew
K Shim
A Gionis
J Gehrke
E Torng
R Agrawal
A Goldberg
A Goel
C Stein
L Zhang
M Mitzenmacher
D Eppstein
J Kleinberg
P Kopke
K Kedem
O Devillers
M Golin
S Schirra
G Shpitalnik
M Katz
A Efrat
B Berger
T Yan
S Ketchpel
A Crespo
B Cooper
J Yang
L Page
H Gupta
Y Sagiv
H Garcia-Molina
O Buyukokten
W Labio
S Melnik
J Widom
J Ullman
K Shoens
A Tomasic
L Raschid
P Buneman
C Chang
S Brin
S Nestorov
T Winograd
D Suciu
R Wilson
M Gavrilov
S Chaudhuri
S Phillips
A Gupta
D Shmoys
S Leonardi
R Gawlick
S Irani
M Fischer
A Tamir
B Thomas
L Cowen
F Chudak
A Borodin
A Fiat
Y Ravid
N Lynch
R Segala
B Lampson
M Segal
C Varghese
A Glozman
A Pogosyants
S Garland
N Delisle
R De Prisco
A Fekete
C Heitmeyer
H Jensen
E Stark
A Shvartsman
D Touitou
R Khazan
E Dolginova
I Keidar
R Friedman
D Foster
G Calinescu
N Shavit
H Karloff
J Drake
O Goldreich
M Hochbaum
A Bar-Noy
A Tomkins
A Schulz
V Vazirani
S Ben-David
D Gibson
N Campbell
H Gabow
R Alur
H Attiya
A Herzberg
R Meir
M Skutella
Y Dinitz
N Tishby
T Tamir
S Fine
G Taubenfeld
E Gafni
F Rendl
S Zhukov
A Fouren
A Gorbach
E Dagan
T Djerassi-Shi
R Ravi
P Indyk
D Hsu
J Feigenbaum
J Latombe
F Ergun
R Motwani
D Aingworth
Z Chen
R Kumar
J Sethi
S Kumar
W Hasan
J Pedersen
S Fujiwara
H Schutze
J Hellerstein
R Impagliazzo
B Cook
C Papadimitriou
V Sarkar
D Papadias
N Moni
M Grigni
E Koutsoupias
S Reyen
T Carson
V Mirelli
X Deng
T Kameda
S Felperin
M Blum
K Kamiya
M Rotcheisen
C Mogensen
P Todd
S Tsur
A Rao
S Bullock
M Shah
J Davis
K French
W Shen
L Russell
B Kero
D Miller
J Josephson
M Elsass
L Anderson

# Computational Efficiency

- Parameter estimation can scale from O(ne) to $O(2^{n(n-1)})$
- Algorithms and data structures for efficient computation
  - H-index for change-score statistics
  - Nets and net-trees
  - Efficient clique-finding algorithms

# Example

- $G = \{V, E\}$

  $V$ = set of $n$ nodes

  $E$ = set of directed binary edges

- Exponential random graph (ERG) model

  $$P(G \mid \theta) = f(G\,;\,\theta) \,/\, \textit{normalization constant}$$

  The normalization constant = sum over all possible graphs

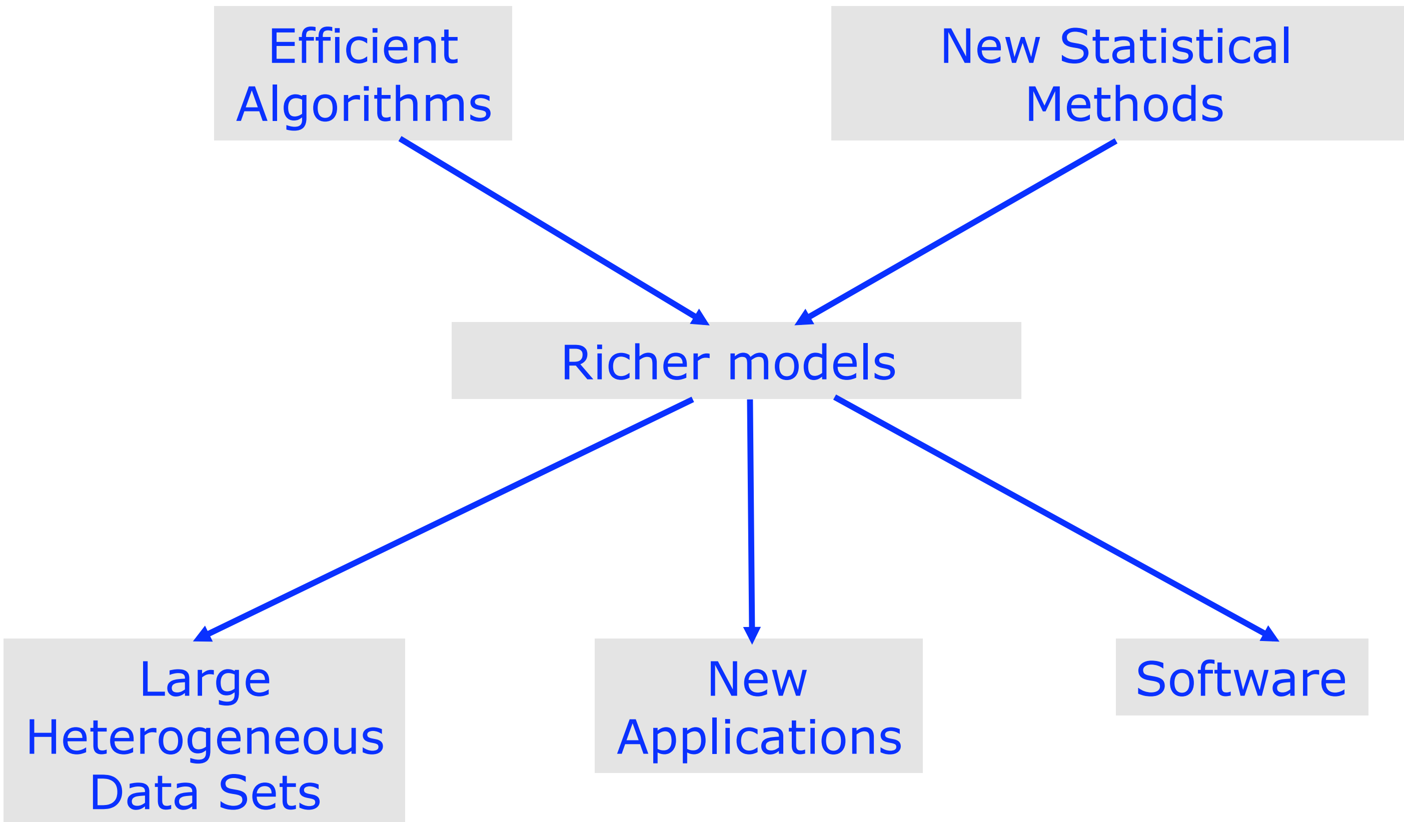
  How many graphs? $2^{n(n-1)}$

  e.g., $n = 50$, we have $2^{2450} \sim 10^{245}$ graphs to sum over

# Key Themes of our MURI Project

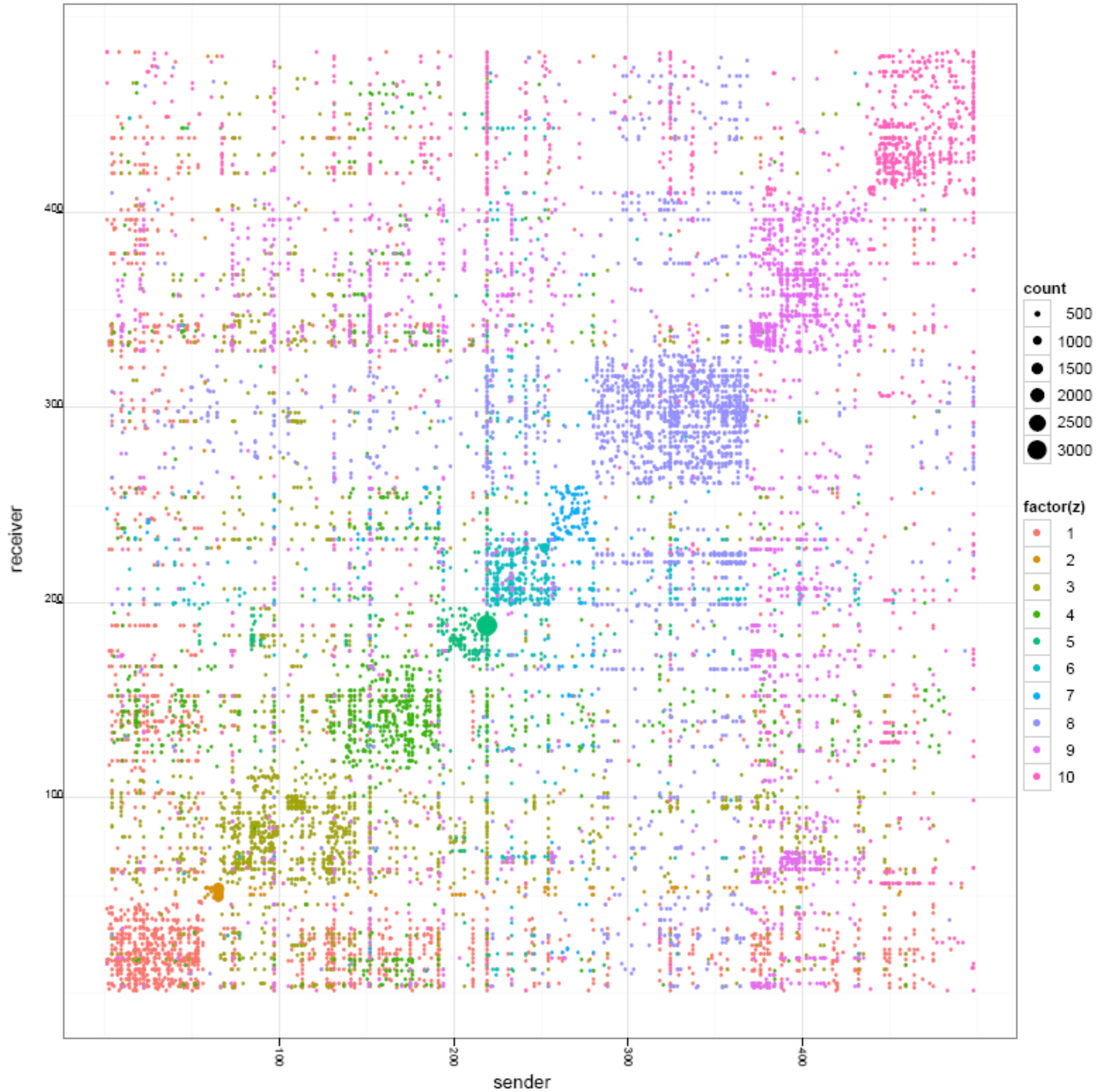
- Research on new statistical estimation techniques and models
  - e.g., principles of modeling and predicting networks over time
- Faster algorithms
  - e.g., efficient data structures and algorithms for very large data sets
- New algorithms for heterogeneous network data
  - Incorporating spatial information, text, other covariates
- Software
  - Make network inference software publicly-available (in R)

# Key Themes of our MURI Project



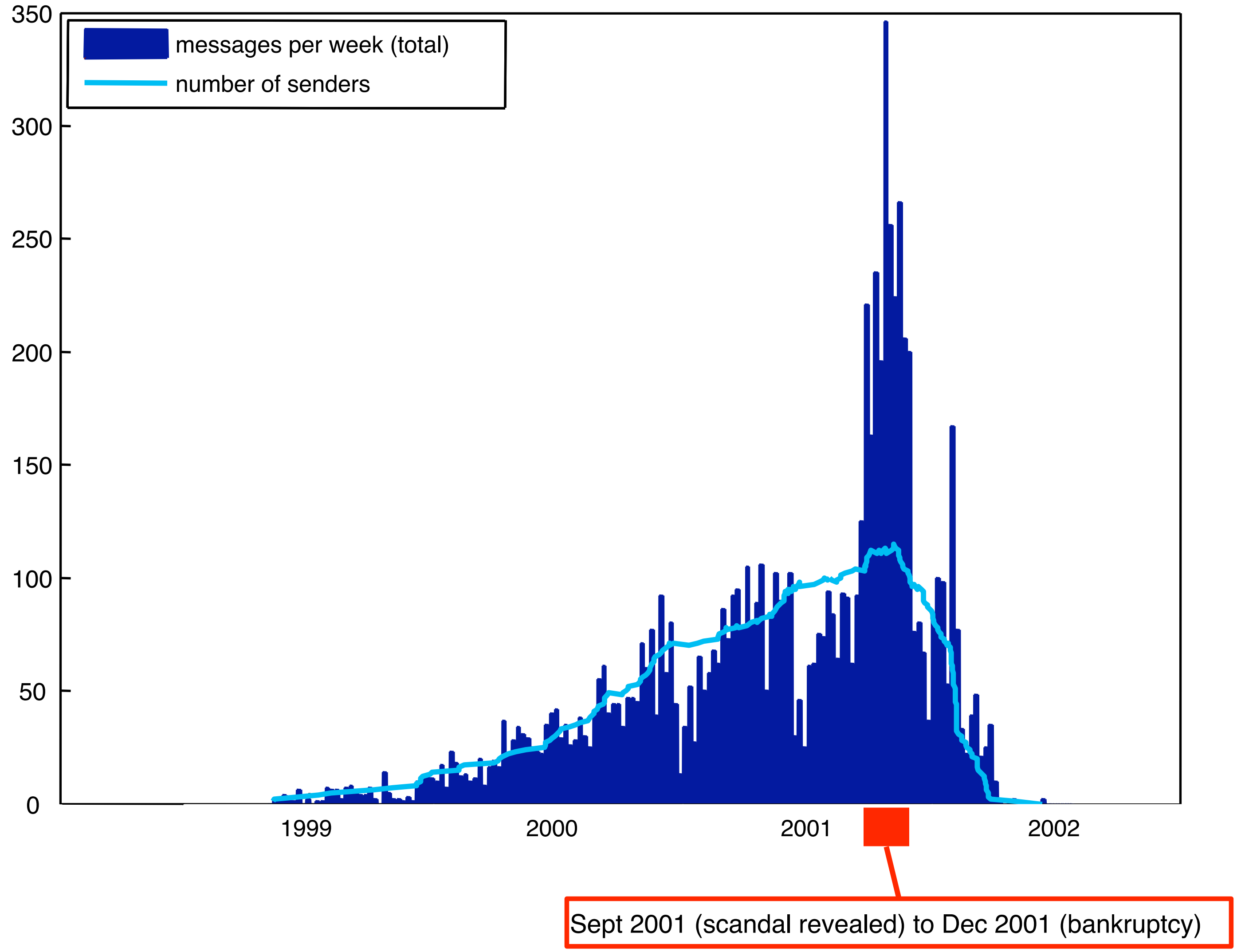
# Complexities of Real Network Data

- Data types
  - Actors and ties
  - Covariates
  - Temporal events
  - Spatial
  - Text
- Structure
  - Hierarchies and clusters
- Measurement issues
  - Sampling
  - Missing data

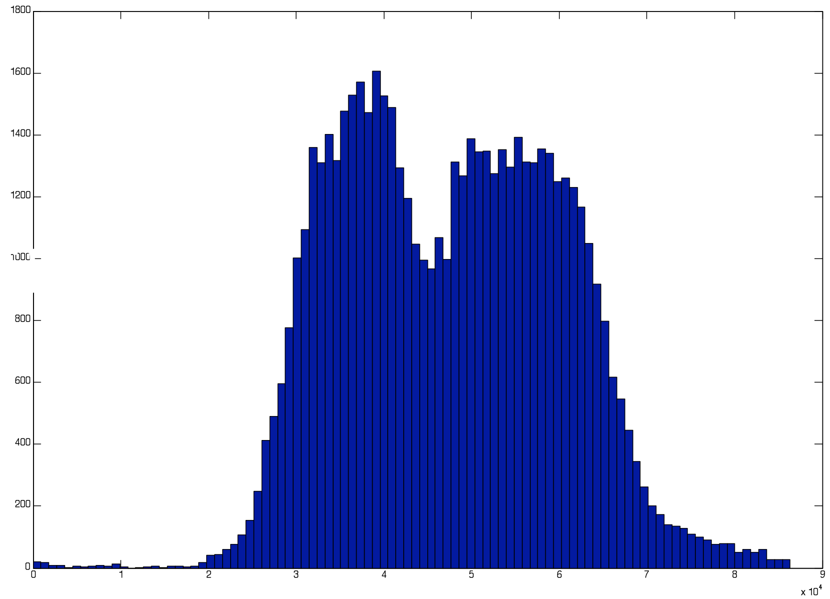


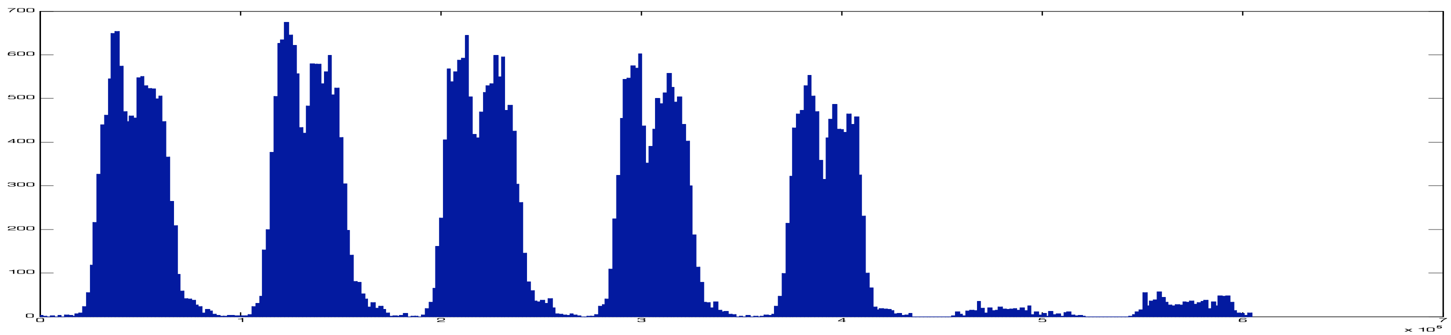

DuBois and Smyth, 2010

# Enron Email Data



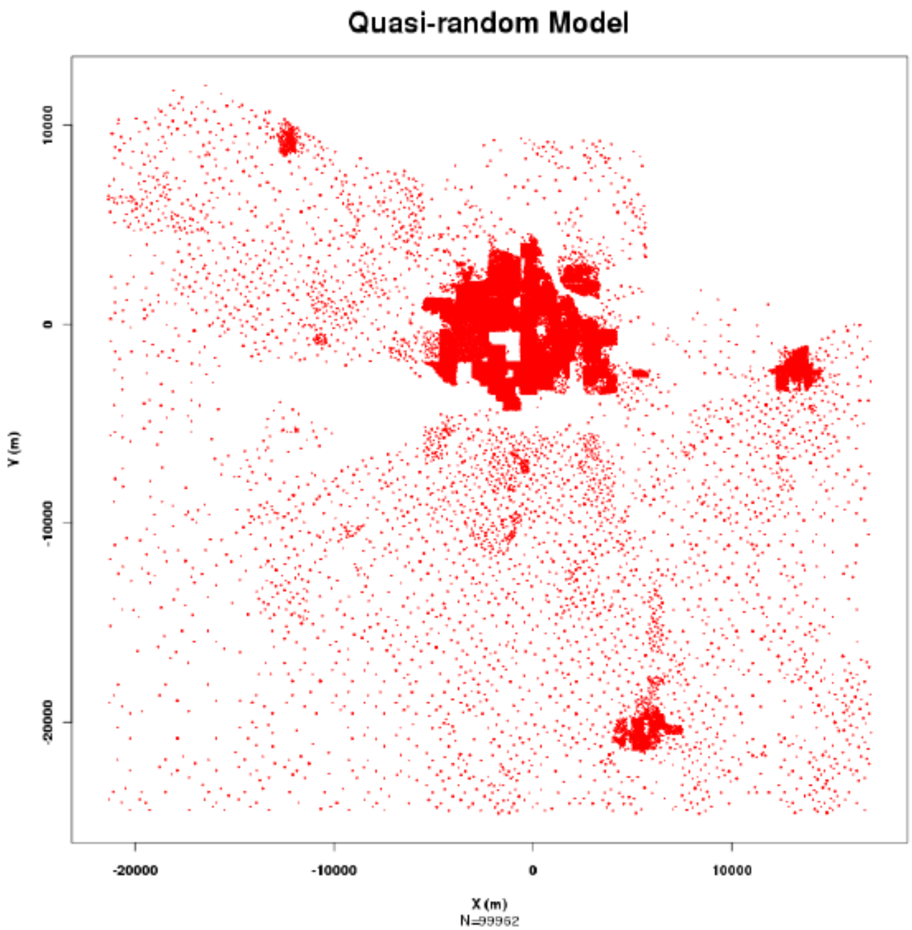
# Daily and weekly variation

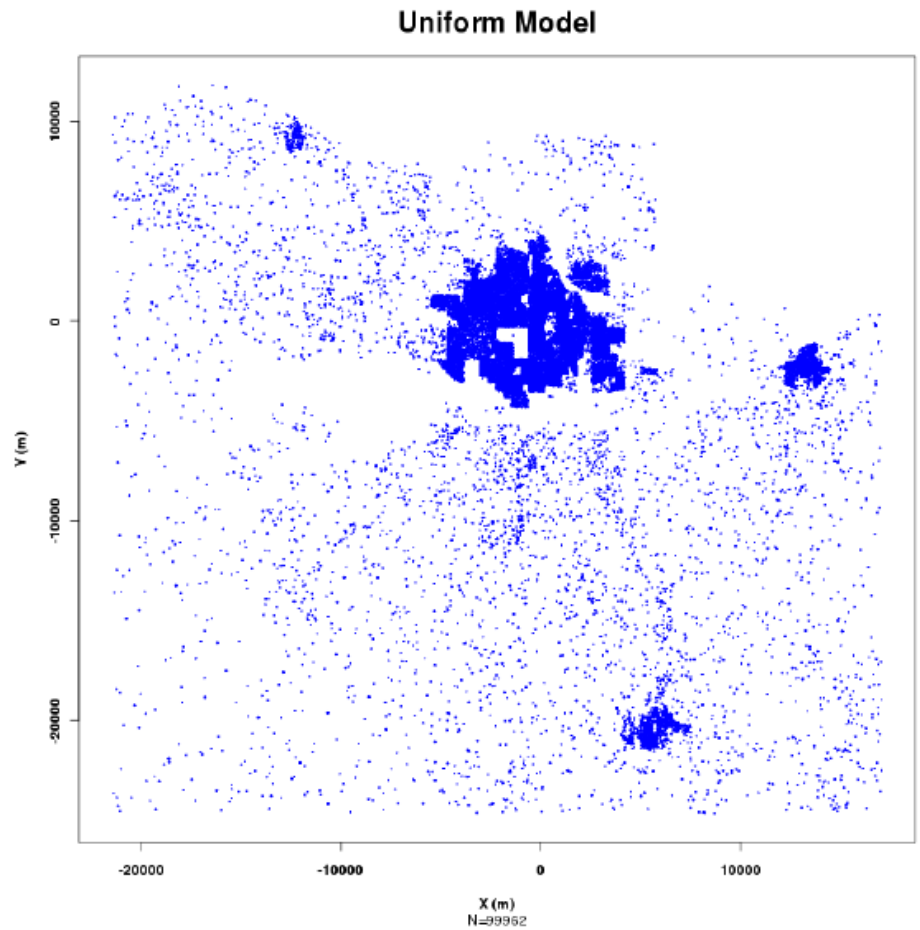


# Spatially-Embedded Network Data

Butts, Acton, Almquist, 2009



# Missing Data

Handcock and Gile, 2008

$Y =$

| | A | B | C | D |
|---|---|---|---|---|
| A | - | 1 | 0 | 0 |
| B | 0 | - | 1 | 1 |
| C | 0 | 0 | - | 0 |
| D | 1 | 1 | 1 | - |

$Y_{obs} =$

| | A | B | C | D |
|---|---|---|---|---|
| A | - | ? | ? | ? |
| B | ? | - | ? | ? |
| C | 0 | 0 | - | 0 |
| D | 1 | 1 | 1 | - |

# Statistical Modeling Frameworks

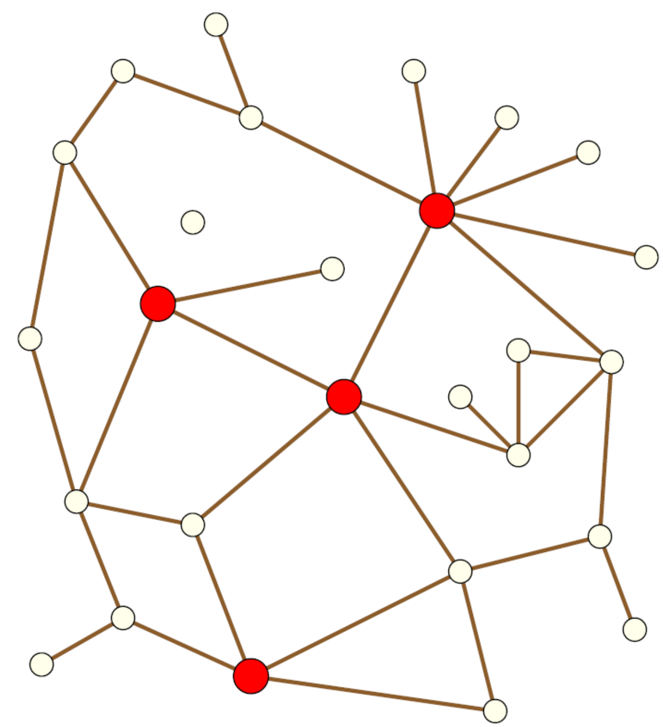
- Exponential random graph models
- Latent-space models
- Relational event models

All 3 frameworks are related – many talks today will touch on at least one of these frameworks

# h-index Data Structures

Eppstein and Spiro, 2009

h-index = maximum number such that
h vertices each have at least h neighbors



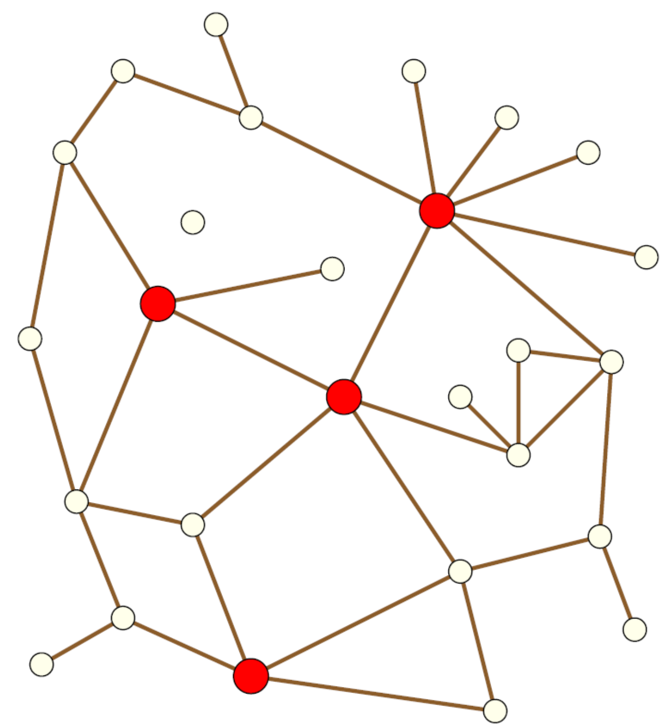
# h-index Data Structures

Eppstein and Spiro, 2009

h-index = maximum number such that
h vertices each have at least h neighbors

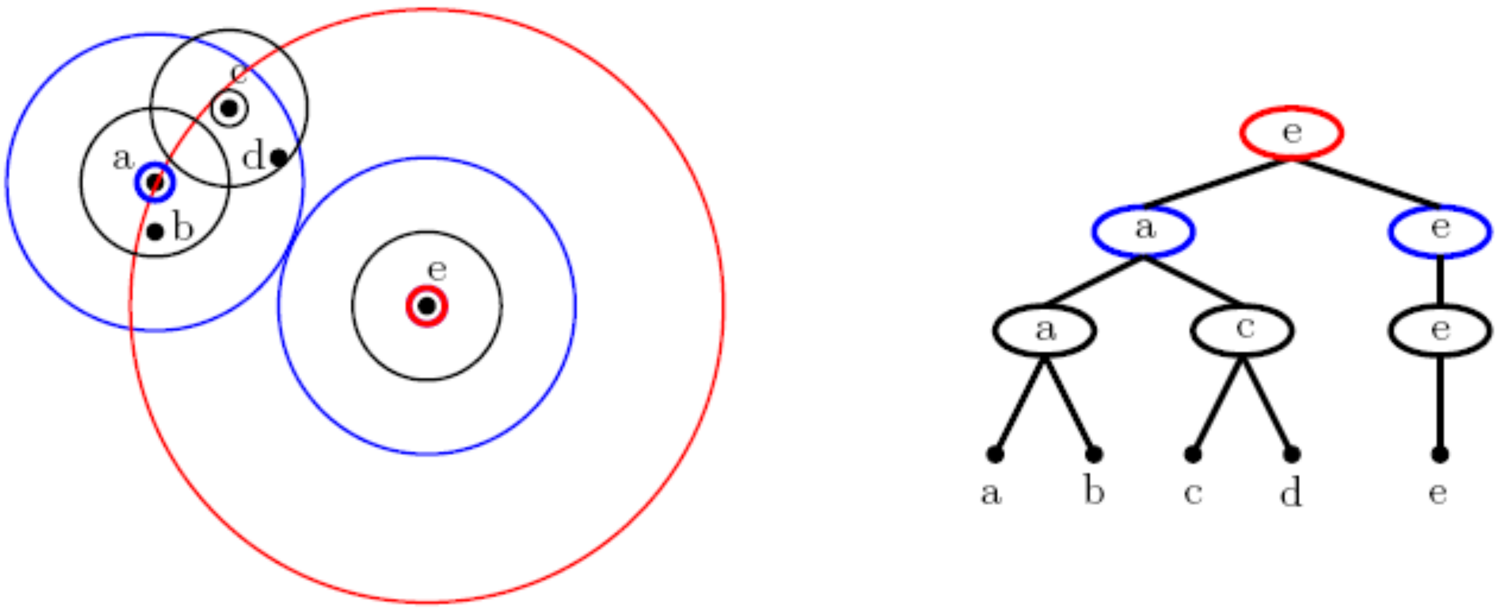
H = set of h high-degree vertices
L = remaining vertices

Can use H/L partitioning to efficiently compute and track graph statistics in statistical estimation algorithms



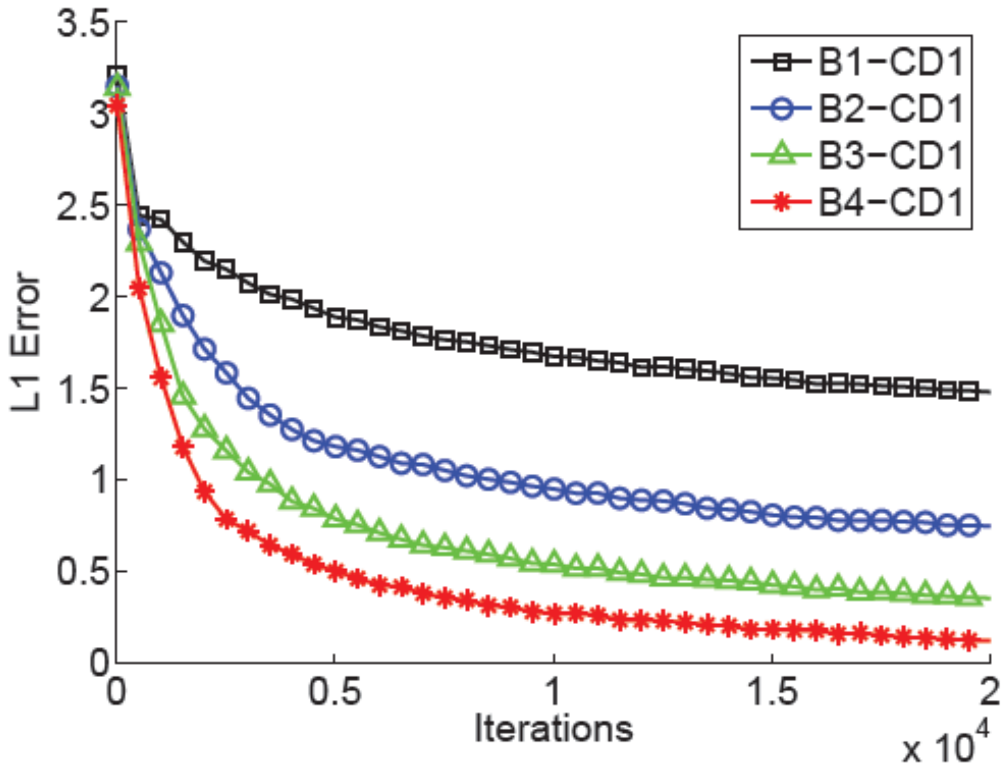
# Nets and Net Trees

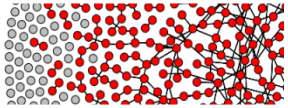
Cho, Mount, Park, 2009



# Fast Sampling Methods

Asuncion et al, 2009





# Evaluation and Prediction

- Evaluate algorithms on large real-world data sets
  - Disaster response
    - Katrina communication networks, World Trade Center disaster response data
  - Networks of documents
    - Political blogs, Wikipedia
  - Social activities on the Web
    - Twitter data, Facebook networks, email communication networks
  - International relations
  - … and more

- Evaluation metrics
  - Computational efficiency
  - Goodness of fit and predictive accuracy

# ONR Interests

(adapted from presentation/discussion in Nov 2008 by Martin Kruger, ONR)

- How does one select the features in an ERG model?
- How can one uniquely characterize a person or a network?
- Can a statistical model (e.g., a relational event model) be used to characterize the trajectory of an individual or a network over time?
- Can one do "activity recognition" in a network?
- Can one model the effect of exogenous changes (e.g., "shocks") to a network over time?
- Importance of understanding social science aspect of network modeling: what are human motivations and goals driving network behavior?

# Morning Session I

| | |
|---|---|
| 9:00 | Introduction and review of project progress<br>Padhraic Smyth (UCI) |
| 9:20 | Implementation issues for latent-space embeddings<br>David Mount (U Maryland) |
| 9:40 | Near-optimal fixed parameter tractability of the Bron-Kerbosch algorithm for maximal cliques<br>Darren Strash (UCI) |
| 10:10 | Methods for analysis of behavioral time-use data<br>Chris Marcum (UCI) |
| 10:30 | BREAK |

# Morning Session II

10:50 Mixture models for event-based network data
Chris DuBois (UCI)

11:10 Static and dynamic robustness in emergency-phase communication networks
Sean Fitzhugh (UCI)

11:30 Bernoulli graph bounds for general random graph models
Carter Butts (UCI)

LUNCH BREAK
12:00 Lunch for ALL meeting participants in 4011

# Afternoon Session I

1:30 Social network analysis of Twitter data
Emma Spiro (UCI)

2:00 Logistic network regression for scalable analysis of dynamic relational data: an overview and case study
Zack Almquist (UCI)

2:20 Latent feature models for network data over time
Jimmy Foulds (UCI)

2:40 New directions in greedy routing on social networks: the membership dimension
Lowell Trott (UCI)

3:10 BREAK

# Afternoon Session II

3:30 Bias-adjusted maximum likelihood estimation methods
Dave Hunter (Penn State)

3:50 Composite likelihood methods for network estimation
Arthur Asuncion (UCI)

4:10 Discussion and Wrap-up
- AHM meeting in November/December
- collaborative activities
- action items

4:30 ADJOURN

# Logistics

- Meals
  - Lunch in this room, 12 noon
  - Refreshment breaks at 10:30 and 3:10

- Wireless
  - Should be able to get 24-hour guest access from UCI network

- Slides will be posted online on the project Web site
  www.datalab.uci.edu/muri

- **Questions and discussion are encouraged during talks!**

# Questions?

# Preprints

R.M. Hummel, M.S. Handcock, D.R. Hunter, A steplength algorithm for fitting ERGMs, submitted, 2009

C. T. Butts, A behavioral micro-foundation for cross-sectional network models, preprint, 2009

C. T. Butts, A perfect sampling method for exponential random graph models, preprint, 2009

A. Asuncion and M. Goodrich, Turning privacy leaks into floods: Surreptitious discovery of Facebook friendships and other sensitive binary attribute vectors, submitted, 2009.

A. Asuncion, Q. Liu, A. Ihler, P. Smyth, Learning with blocks: composite likelihood and contrastive divergence, submitted, 2009.

# Tasks

A: Fast network estimation algorithms
    Eppstein, Butts

B: Spatial representations and network data
    Goodrich, Eppstein, Mount

C: Advanced network estimation techniques
    Handcock, Hunter

D: Scalable methods for relational events
    Butts

E: Network models with text data
    Smyth

F: Software for network inference and prediction
    Hunter

# Estimation Algorithms

- We want P(parameters | data)
- Exact algorithms are rare
- Approximate search
  - E.g., Markov chain Monte Carlo
- Exact solution of simpler objective function
  - E.g., pseudolikelihood v. likelihood

# Collaboration Network