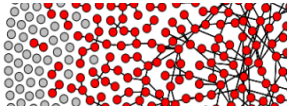
Scalable Methods for the Analysis of Network-Based Data

MURI Project: University of California, Irvine

Annual Review Meeting

December 8th 2009

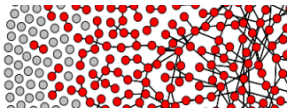
Principal Investigator: Padhraic Smyth



Today's Meeting

- Goals
  - Review our research progress
  - Feedback from project sponsors (ONR)

- Format
  - Introduction
  - Tutorial talks
  - Research updates from each PI
  - Poster session by graduate students
  - Discussion and feedback



Project Dates

- Project Timeline
  - Start date: May 1 2008
  - End date: April 30 2011/2013

- Meetings
  - Kickoff Meeting, November 2008
  - Working Meeting, April 2009
  - Working Meeting, August 2009
  - Annual Review, December 2009

[meeting slides online at www.datalab.uci.edu/muri]

MURI Investigators

Padhraic Smyth
UCI

David Eppstein
UCI

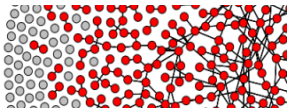
Carter Butts
UCI

Michael Goodrich
UCI

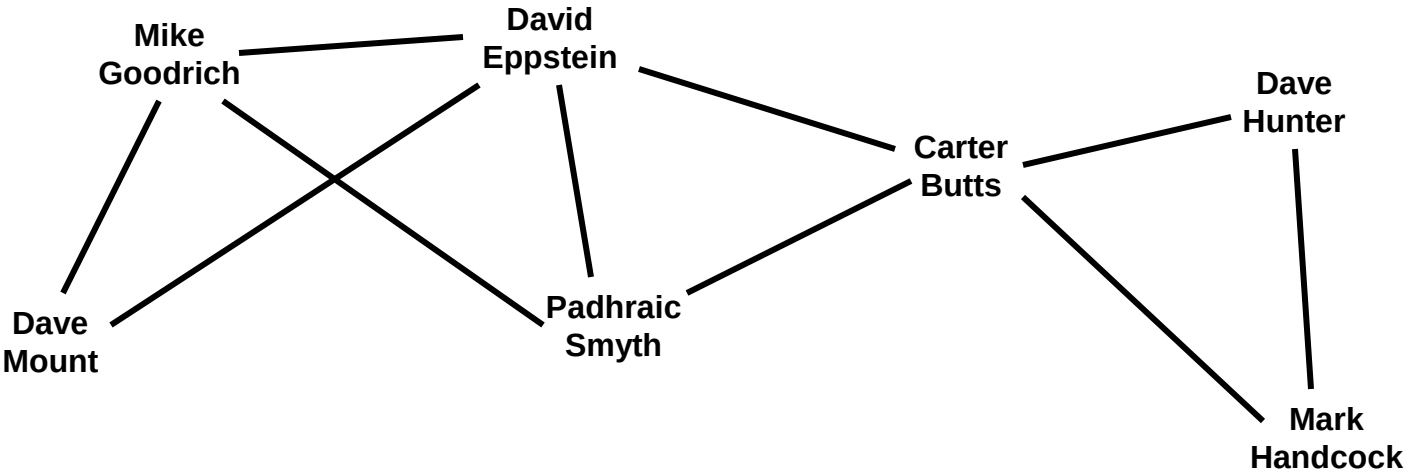
Mark Handcock
U Washington

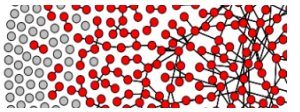
Dave Mount
U Maryland

Dave Hunter
Penn State

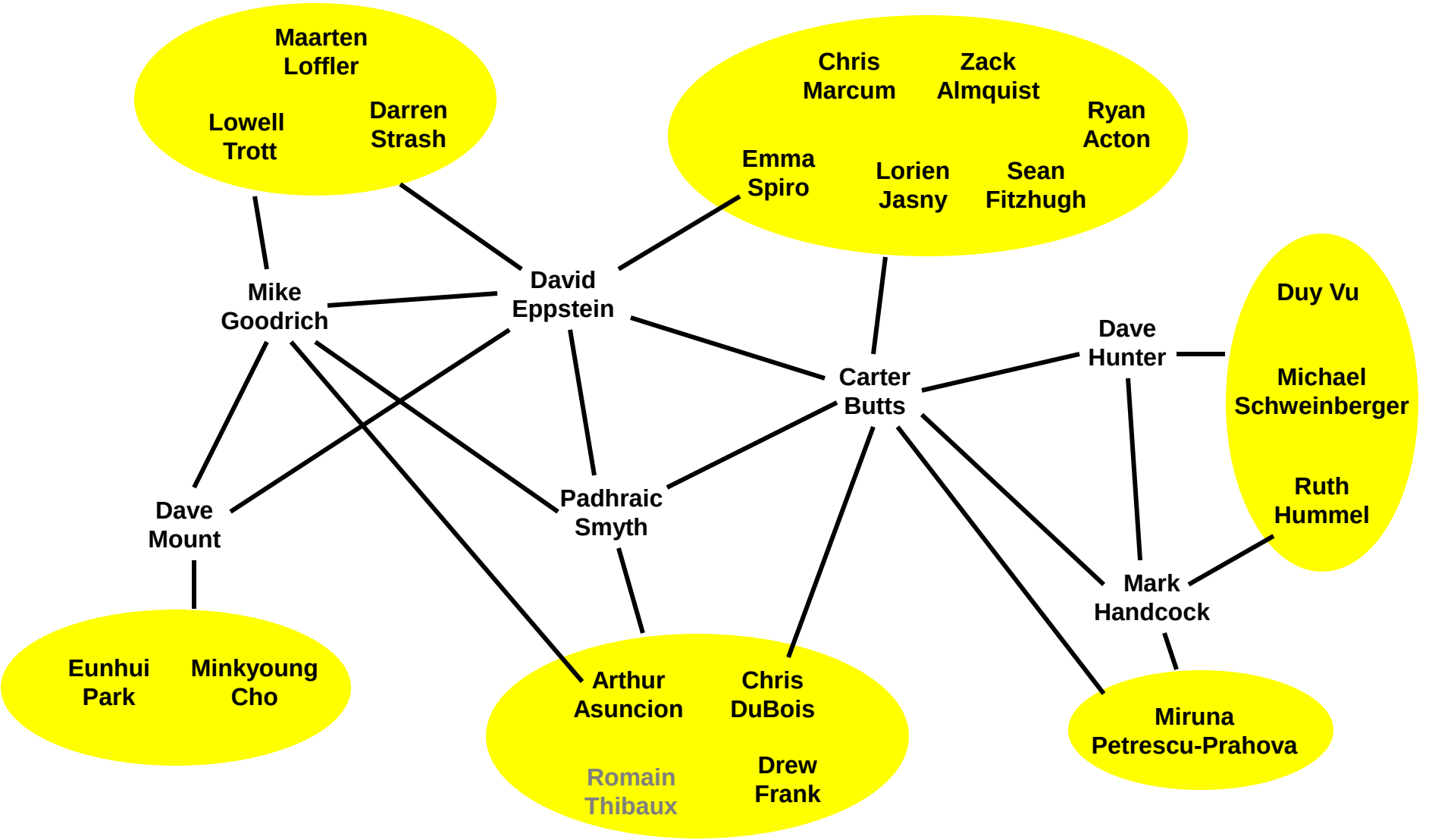


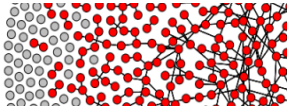
Collaboration Network





Collaboration Network





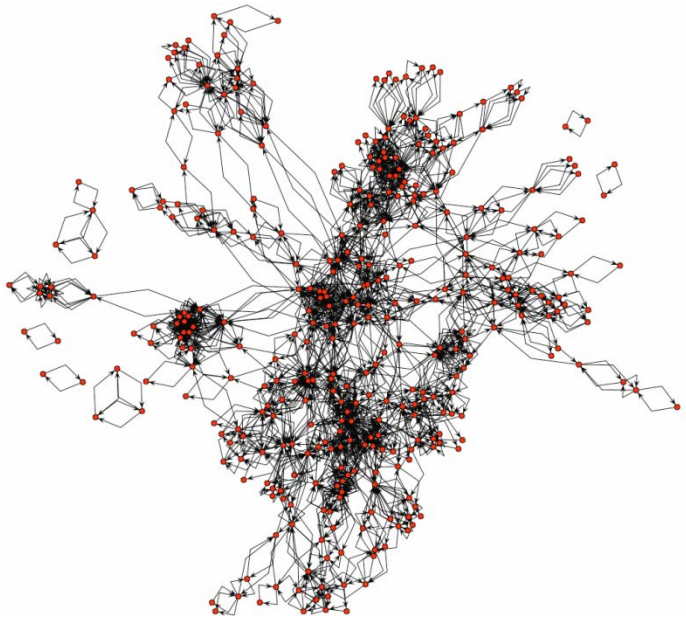
Data

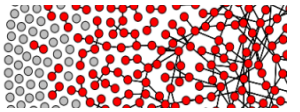
Statistical Models

Scalable Algorithms

Evaluation

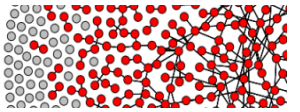
Software and Applications





Limitations of Existing Methods

- Computational intractability
  - Current statistical network modeling algorithms can scale exponentially in the number of nodes *N*
- Network data over time
  - Relatively little work on statistical models for dynamic network data
- Heterogeneous data
  - e.g., few techniques for incorporating text, spatial information, etc, into network models



Example

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

  V = set of N nodes

  E = set of directed binary edges

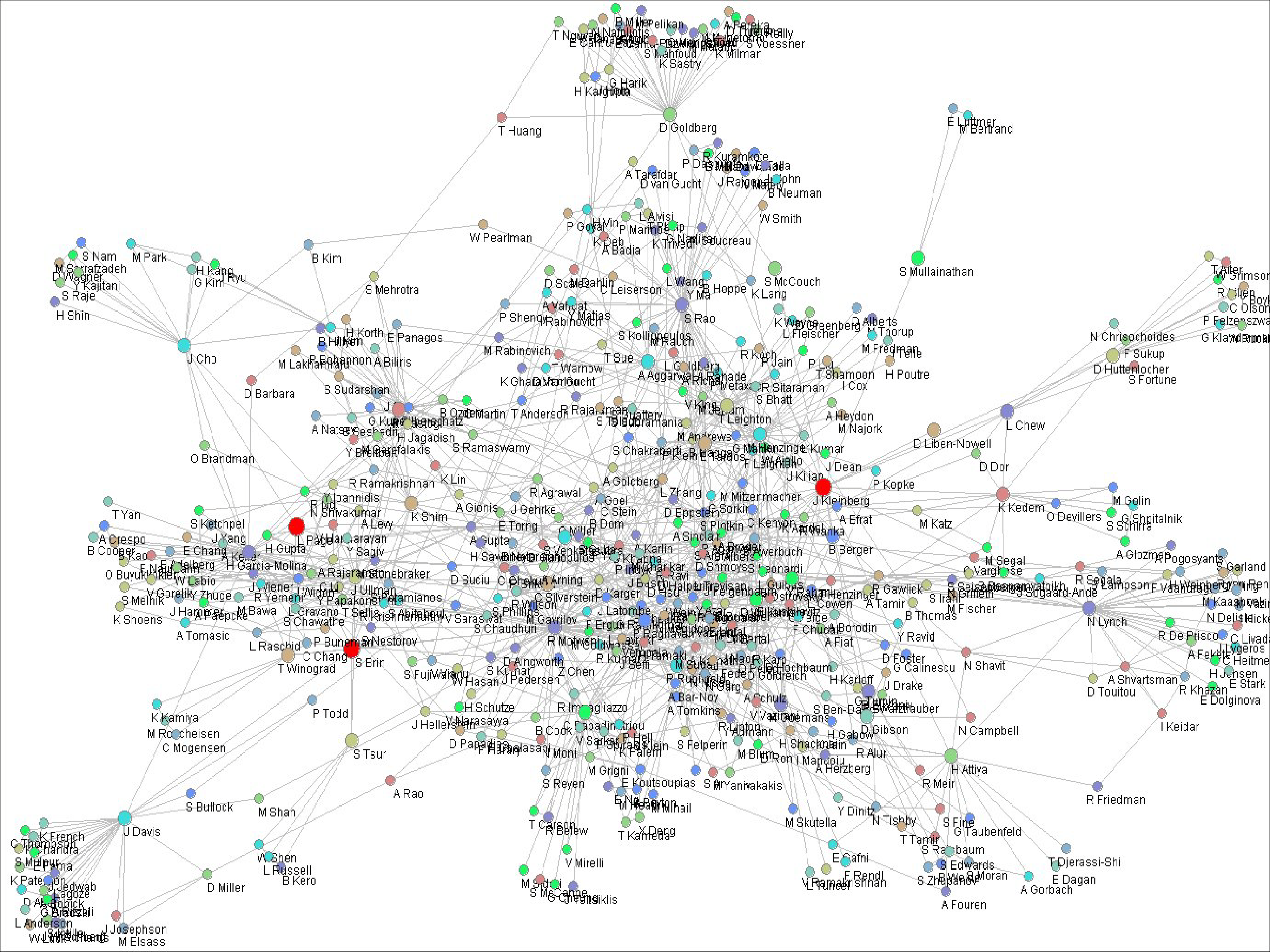
- Exponential random graph (ERG) model

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

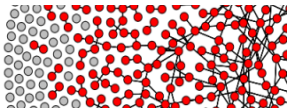
The normalization constant = sum over all possible graphs

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

e.g., $N = 20$, we have $2^{380} \sim 10^{38}$ graphs to sum over

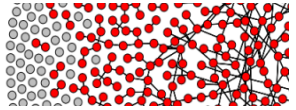


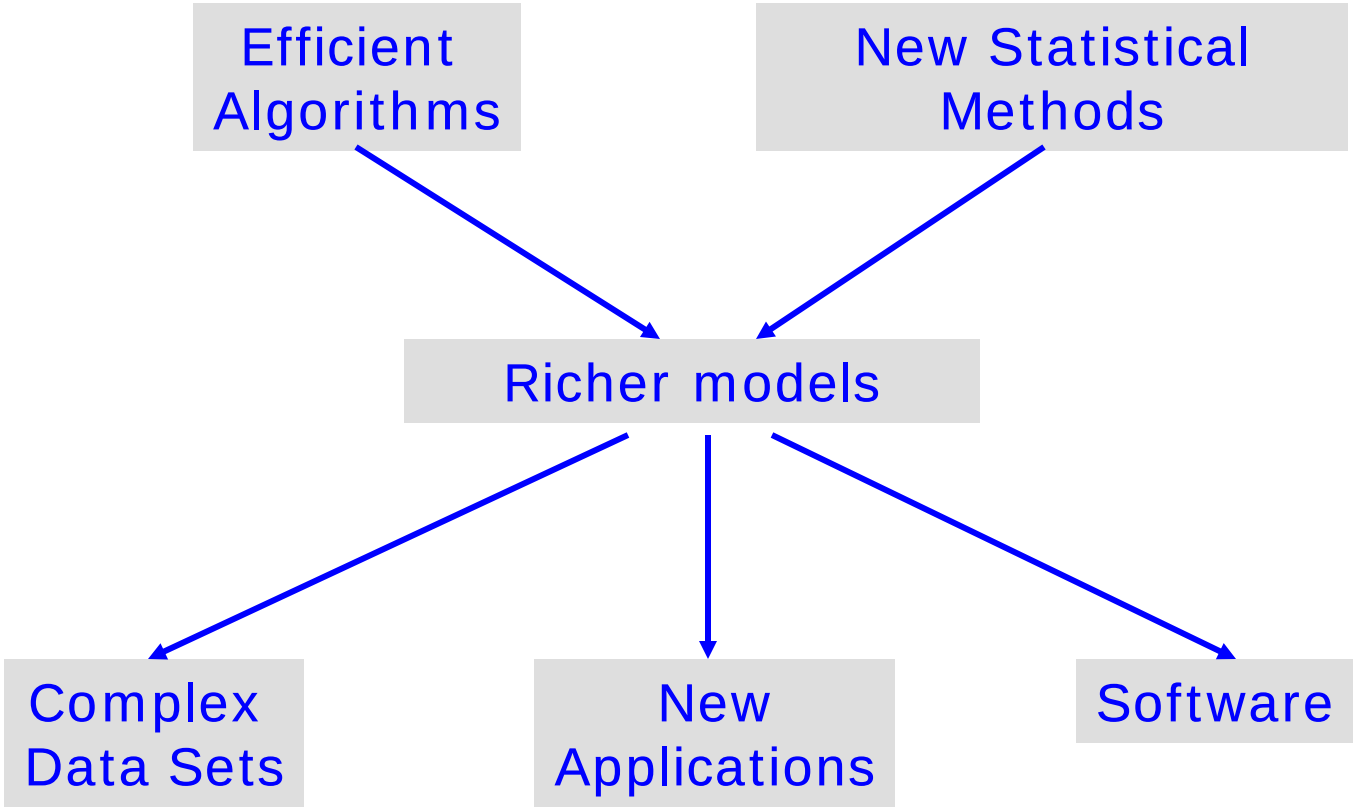
T Huang
D Goldberg
K Sastry
S Mahfoud
K Milman
S Voessner
G Harik
E Luttmer
M Bertrand
A Tarafdar
D van Gucht
B Neuman
W Smith
S Mullainathan
W Pearlman
B Kim
M Park
S Nam
M Sarrafzadeh
D Wagner
Y Kajitani
S Raje
H Shin
H Kang
S Kim
J Ryu
J Cho
S Mehrotra
S McCouch
K Lang
S Rao
P Shenoy
T Alter
W Grimson
C Olson
P Felzenszwalb
N Chrisochoides
F Sukup
D Huttenlocher
S Fortune
H Korth
E Panagos
A Biliris
M Lakhamraju
S Sudarshan
D Barbara
M Rabinovich
T Warnow
T Suel
A Aggarwal
L Fleischer
D Alberts
M Thorup
M Fredman
H Poutre
T Shamoon
I Cox
S Bhatt
A Heydon
M Najork
L Chew
D Liben-Nowell
D Dor
T Anderson
H Jagadish
S Ramaswamy
O Brandman
T Leighton
M Andrews
S Chakrabarti
J Kleinberg
J Dean
J Kilian
P Kopke
K Kedem
D Eppstein
C Stein
M Mitzenmacher
R Ramakrishnan
K Lin
R Agrawal
A Gionis
J Gehrke
A Goldberg
L Zhang
K Shim
S Plotkin
C Kenyon
A Efrat
M Katz
O Devillers
M Golin
S Schirra
G Shpitalnik
T Yan
S Ketchpel
J Yang
A Crespo
B Cooper
H Gupta
H Garcia-Molina
E Chang
A Gupta
E Torng
B Dom
C Miller
A Karlin
B Berger
M Segal
A Glozman
A Pogosyan
S Garland
D Suciu
N Lynch
O Buyukkokten
W Labio
S Melnik
K Shoens
A Paepcke
A Tomasic
L Raschid
T Winograd
C Chang
P Buneman
S Brin
S Nestorov
M Bawa
S Chawathe
J Ullman
M Stonebraker
S Abiteboul
R Krishnamurthy
S Chaudhuri
M Gavrilov
R Motwani
J Latombe
C Silverstein
D Karger
R Gawlick
A Tamir
M Fischer
B Thomas
A Borodin
F Chudak
A Fiat
Y Ravid
D Foster
G Calinescu
N Shavit
D Touitou
A Shvartsman
R Khazan
E Dolginova
I Keidar
R De Prisco
H Jensen
E Stark
K Kamiya
M Rotcheisen
C Mogensen
P Todd
S Tsur
A Rao
S Fujiwara
W Hasan
J Pedersen
Z Chen
D Aingworth
R Impagliazzo
H Schutze
V Narasayya
J Hellerstein
D Papadias
B Cook
M Grigni
E Koutsoupias
S Reyen
N Moni
S Felperin
M Yannakakis
A Bar-Noy
A Tomkins
A Schulz
Y Aumann
M Blum
D Ron
I Mantoiu
A Herzberg
R Alur
H Gabow
D Gibson
J Drake
H Karloff
N Campbell
H Attiya
R Meir
R Friedman
S Bullock
M Shah
J Davis
K French
C Thompson
D Miller
W Shen
L Russell
B Kero
J Jedwab
L Anderson
J Josephson
M Elsass
T Carson
R Belew
V Mirelli
T Kameda
X Deng
M Mihail
M Skutella
Y Dinitz
N Tishby
T Tamir
S Fine
G Taubenfeld
S Rajsbaum
S Edwards
E Gafni
T Djerassi-Shi
E Dagan
A Gorbach
A Fouren

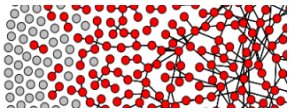


Key Themes of our MURI Project

- Foundational research on new statistical models and methods for social network data
  - e.g., decision-theoretic foundations of social networks
- Efficient estimation algorithms
  - E.g., efficient data structures for very large data sets
- New algorithms for heterogeneous network data
  - Incorporating time, space, text, other covariates
- Software
  - Make network inference software publicly-available (in R)




Efficient Algorithms
New Statistical Methods
Richer models
Complex Data Sets
New Applications
Software



Complex Network Data

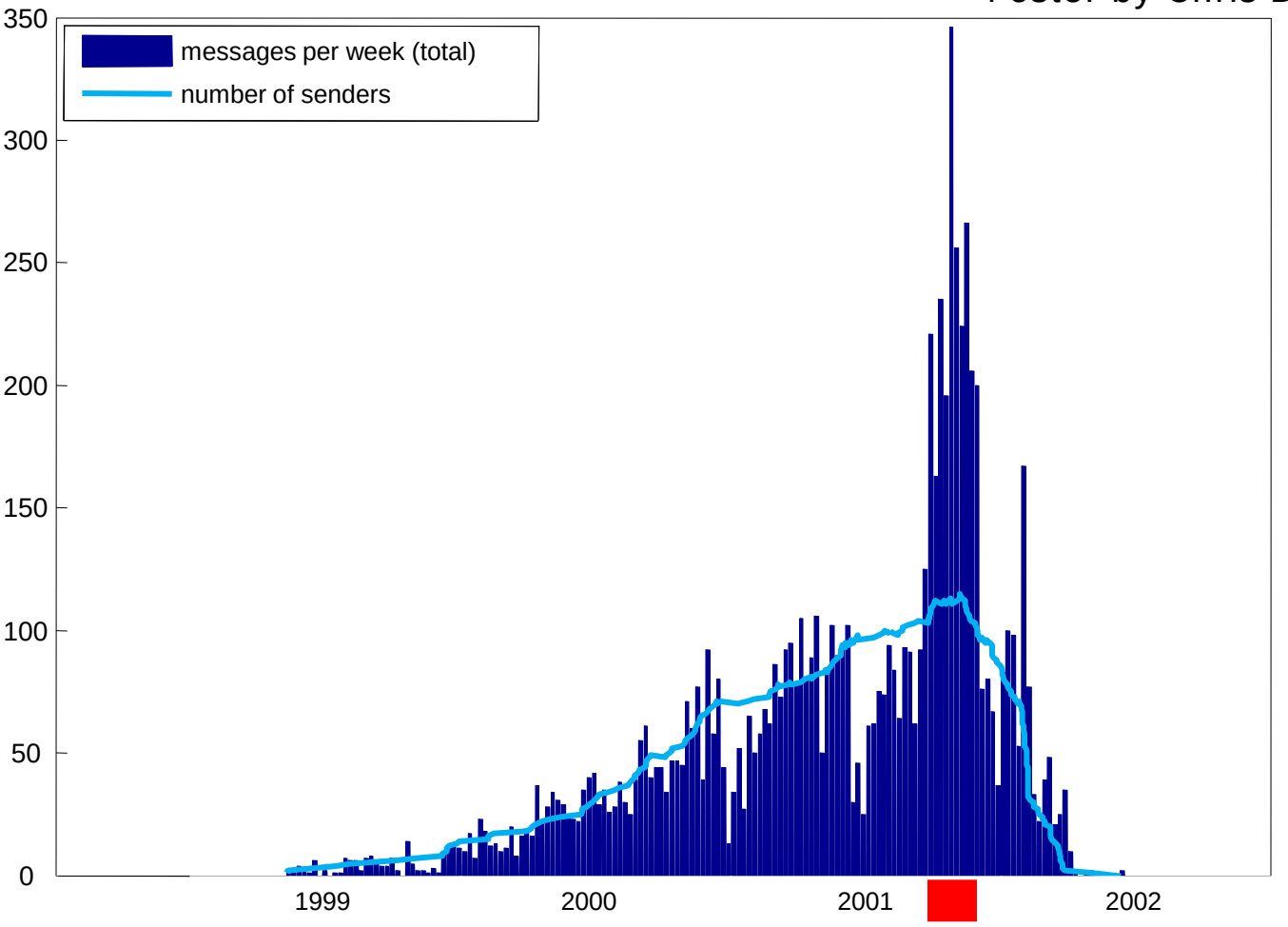
- Data types
  - Actors and ties
  - Temporal events *(Posters by DuBois, Almquist, Jasny, Marcum)*
  - Spatial information *(Poster by Acton)*
  - Text data *(Poster by Asuncion, talk by Smyth)*
  - Actor and tie covariates

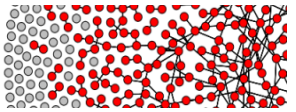
- Structure
  - Hierarchies and clusters
    (Talk by Petrescu-Prahova, Poster by DuBois)

- Measurement issues
  - Sampling
  - Missing data

Enron Email Data

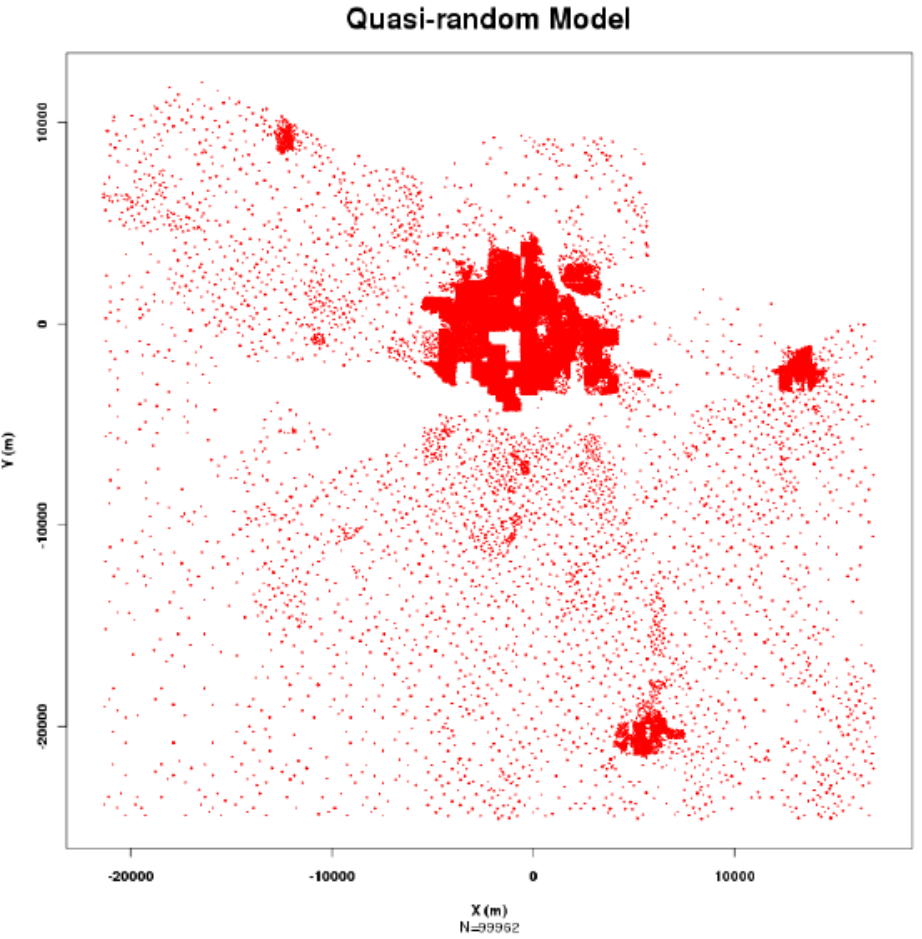
Poster by Chris DuBois

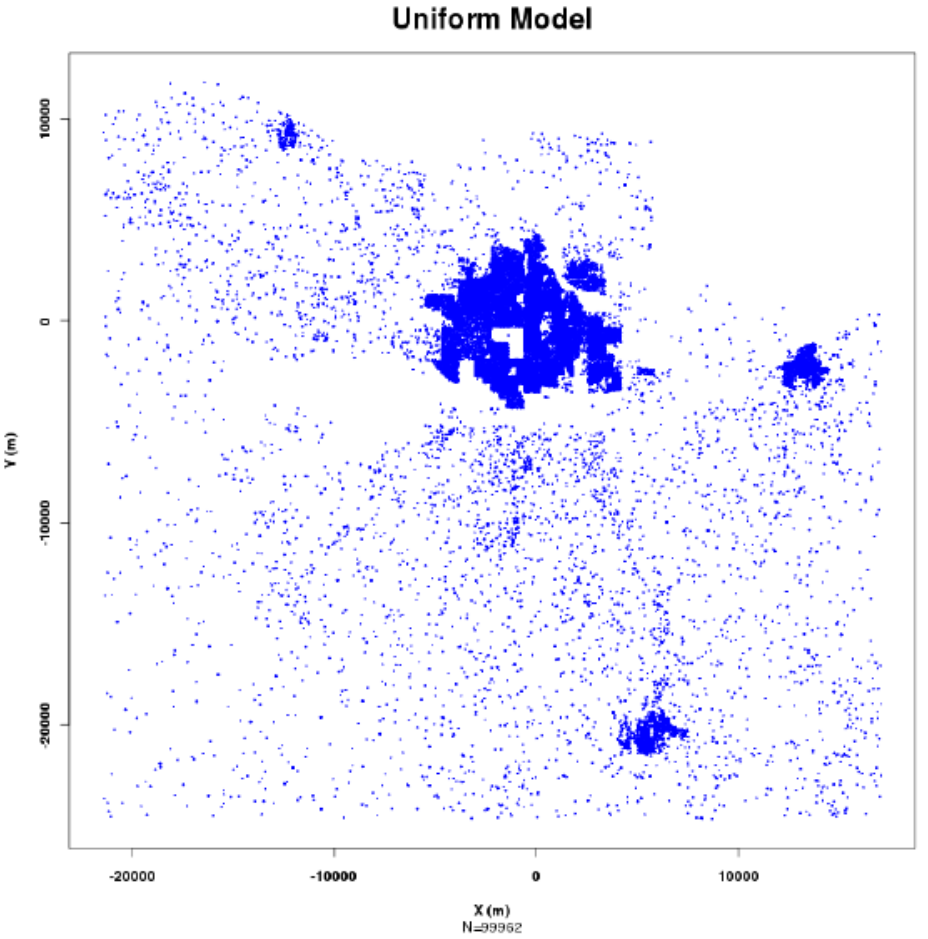


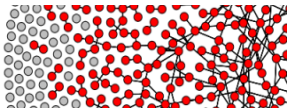


Spatial Network Data

Poster by Ryan Acton







Missing Data

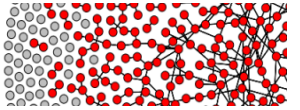
Handcock and Gile, 2008

$Y =$

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

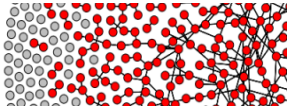
$Y_{obs} =$

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



Statistical Models for Network Data

- Exponential random graph models
  (Talks by Hunter, Eppstein, Petrescu-Prahova)
- Relational event models
  (Posters by Marcum, Jasny)
- Latent-variable models
  (Talks by Mount, Smyth, Petrescu-Prahova)
  (Posters by Asuncion, DuBois)
- Decision-theoretic frameworks for social networks
  (Talk by Butts)

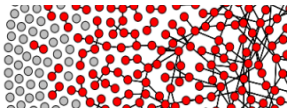


Estimation Algorithms

- We seek P(parameters | data)
- Exact algorithms are rare
- Approximate search
  - E.g., Markov chain Monte Carlo

    (talks by Hunter, poster by Hummel)
- Exact solution of simpler objective function
  - E.g., pseudolikelihood v. likelihood

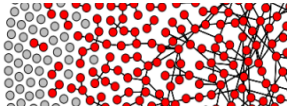
    (talks by Hunter)



Computational Efficiency

- Parameter estimation can scale from O(*Ne*) to O($2^{N(N-1)}$)

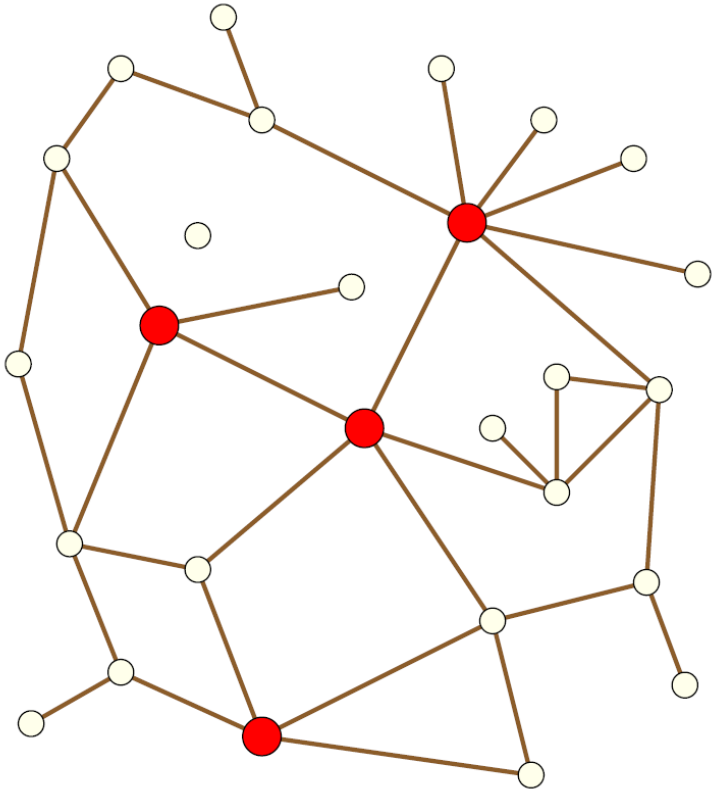
- Data structures for efficient computation:
  - H-index for change-score statistics
    (talk by Eppstein, posters by Spiro and by Trott)
  - Nets and net-trees
    (talk by Mount, poster by Park)
  - Priority range trees
    (poster by Strash)

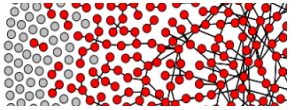


h-index Data Structures

Eppstein and Spiro, 2009

Maximum number of nodes such that h nodes each have at least h neighbors



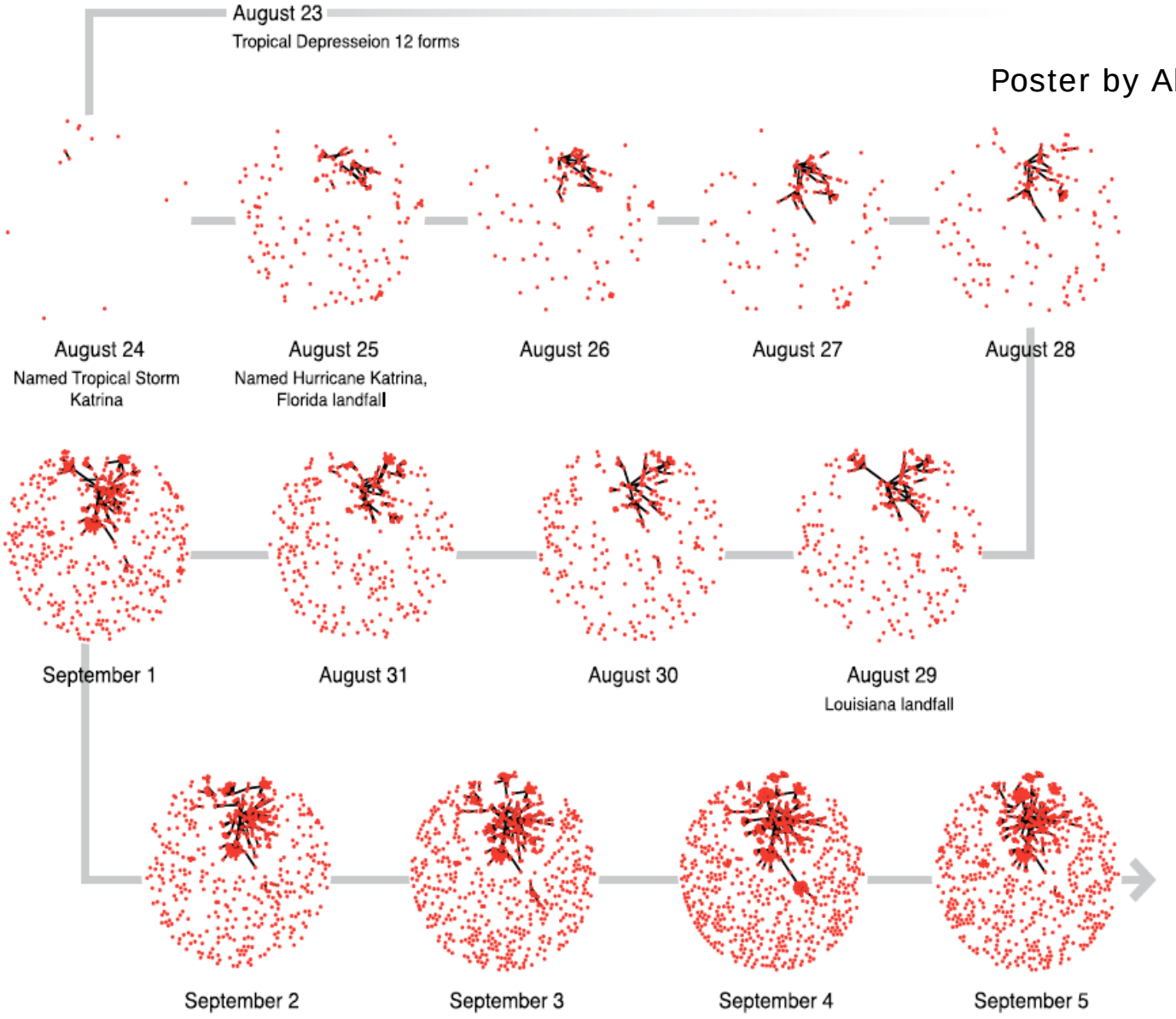


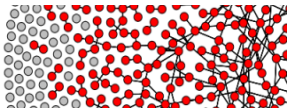
Evaluation and Prediction

- Evaluation on real-world data sets
  - Katrina communication networks
  - World Trade Center disaster response data
  - Political blogs
  - Facebook egonets
  - Facebook UNC
  - Enron email data
  - … and more

- Metrics
  - Assessment of model fit, e.g., BIC criterion
  - Predictive accuracy on test data, e.g., for temporal events

Poster by Almquist





Publications

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

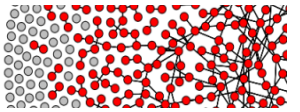
R. Hummel, M. Handcock, D. Hunter, A steplength algorithm for fitting ERGMS, winner of the American Statistical Association (Statistical Computing and Statistical Graphics Section) student paper award, presented at the *ASA Joint Statistical Meeting*, 2009.

D. Eppstein and E. S. Spiro, The h-index of a graph and its application to dynamic subgraph statistics, *Algorithms and Data Structures Symposium*, Banff, Canada, August 2009

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

M. Cho, D. M. Mount, and E. Park, Maintaining nets and net trees under incremental motion, in Proceedings of the 20th International Symposium on Algorithms and Computation, 2009.

M. Gjoka, M. Kurant, C. T. Butts, A. Markopoulou, A walk in Facebook: uniform sampling of users in online social networks, electronic preprint, IEEE Infocom, to appear.



Preprints

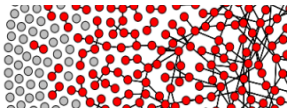
R.M. Hummel, M.S. Handcock, D.R. Hunter, A steplength algorithm for fitting ERGMs, submitted, 2009

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.



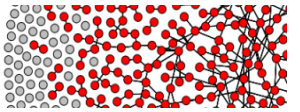
Morning Session I

9:00 Introduction and Overview
Padhraic Smyth, UC Irvine

9:20 Principles of Statistical Network Modeling
Carter Butts, UC Irvine

9:50 Estimation Methods for Statistical Network Modeling
David Hunter, Pennsylvania State University

10:15 Break



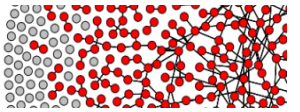
Morning Session II

10:40 Efficient Computation of Change-Graph Scores
David Eppstein, UC Irvine

11:05 Decision-Theoretic Foundations of Statistical Network Models
Carter Butts, UC Irvine

11:30 Privacy Leaks and Floods in Social Networks
Michael Goodrich, UC Irvine

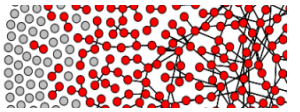
12:00 Break for lunch
- PIs + ONR visitors at the University Club
- Students and postdocs, lunch in 6011



Graduate Student Poster Session

(1:15 to 2:30, in this room, 6011)

Lorien Jasny:	Using Egocentric Relational Event Models to Predict Improvisation
Chris Marcum:	Complex Sequence Terms for Egocentric Relational Event Models
Zack Almquist:	Logistic Model for Network Evolution (Katrina Case)
Sean Fitzhugh:	Effects of Individual and Group-level Properties on World Trade Center Radio Network Robustness
Ryan Acton:	Geographical Models of Large-scale Social Networks
Emma Spiro:	Assessing the Degree h-Index Distribution for Social Networks
Darren Strash:	Priority Range Trees
Lowell Trott:	Extended Dynamic Subgraph Statistics using the h-Index
Chris DuBois:	Stochastic Blockmodels for Network-based Event Data
Arthur Asuncion:	Joint Statistical Models for Text and Social Networks
Ruth Hummel:	A Steplength Algorithm for Fitting ERGMs
Eunhui Park:	A Dynamic Data Structure for Approximate Range Searching

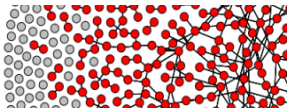


Afternoon Session I

2:30 Algorithms and Data Structures for Embedded Network Data
David Mount, University of Maryland

2:55 Latent Variable Models for Text, Event, and Network Data
Padhraic Smyth, UC Irvine

3:15 COFFEE BREAK



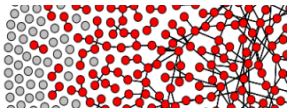
Afternoon Session II

3:40 Scalable Estimation Algorithms for Large Network Data Sets
David Hunter, Pennsylvania State University

4:05 Statistical Inference for Latent Degree-Class Models with Applications to Disaster Networks
Miruna Petrescu-Prahova, University of Washington
and Michael Schweinberger, Pennsylvania State University

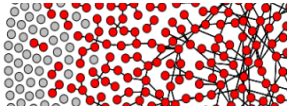
4:30 OPEN DISCUSSION

5:15 ADJOURN

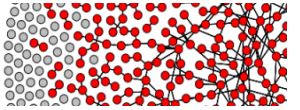


Logistics

- Meals
  - Lunch at University Club - for visitors and PIs
  - Refreshment breaks at 10:30 and 3:15
- Wireless
  - Should be able to get 24-hour guest access from UCI network
- Online Slides and Schedule
  www.datalab.uci.edu/muri
- Reminder to speakers: leave time for questions and discussion!



Questions?



Nets and Net Trees

Cho, Mount, Park, 2009