Scalable Methods for the Analysis of Network-Based Data

MURI Project: University of California, Irvine

Project Meeting

August 25th 2009

Principal Investigator: Padhraic Smyth

Goals for Today's Meeting

- Introductions and brief review of our project
- Technical presentations and discussion
  - MURI-related research, different research groups
  - Important to leave time for questions and discussion
    - 30 minute talks: finish in 25 mins
    - 15 minute talks: finish in 12 mins
  - Goal is to spur discussion and interaction
- End of day
  - Open discussion: research, collaboration
  - Organizational items: date of November meeting
  - Wrap–up and action items

MURI Investigators

Padhraic Smyth
UCI

David Eppstein
UCI

Carter Butts
UCI

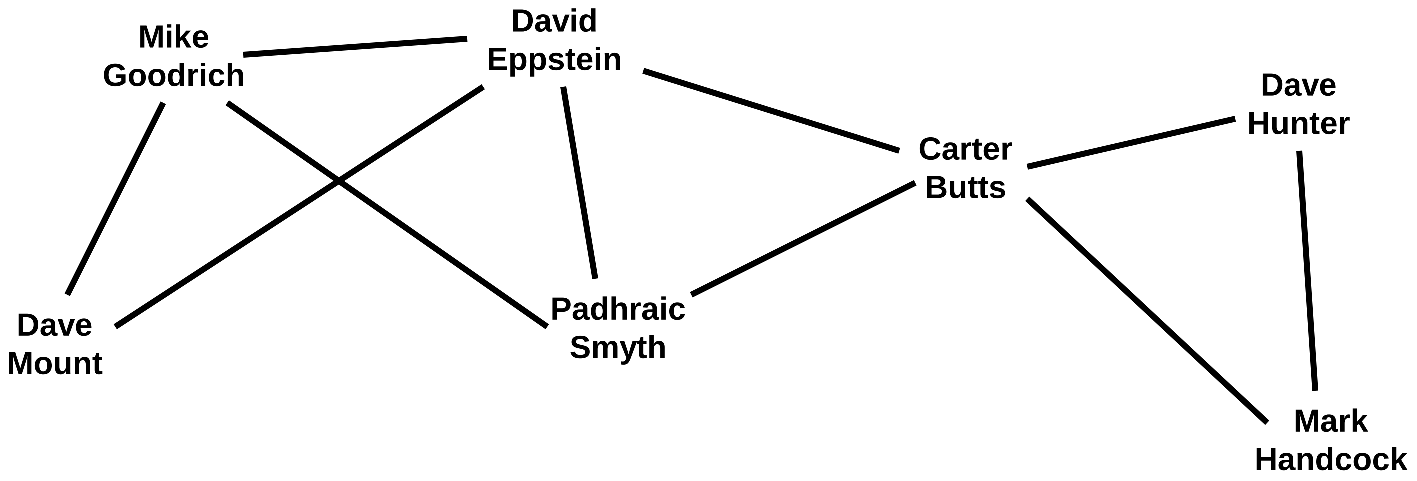
Michael Goodrich
UCI

Mark Handcock
U Washington

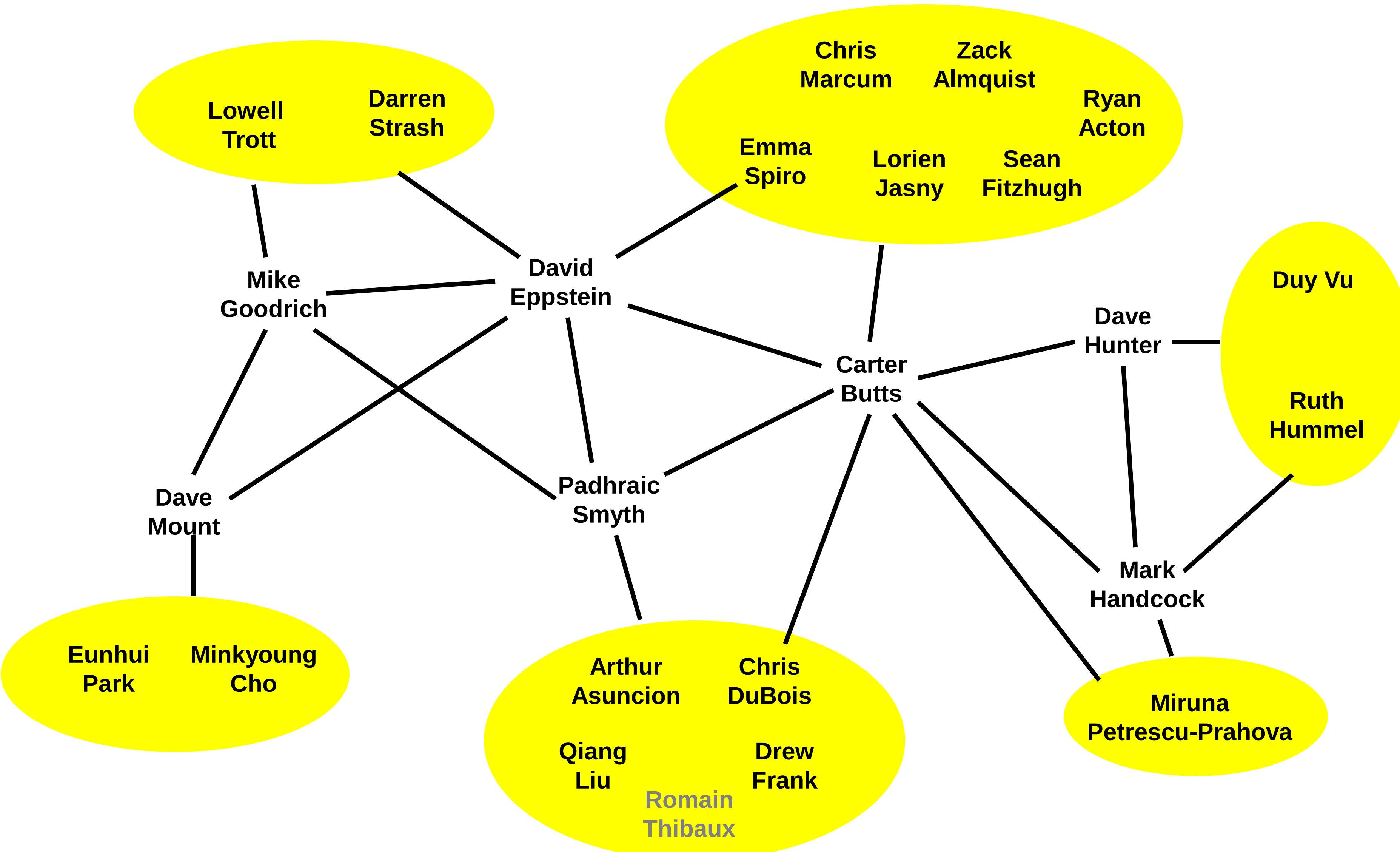
Dave Mount
U Maryland

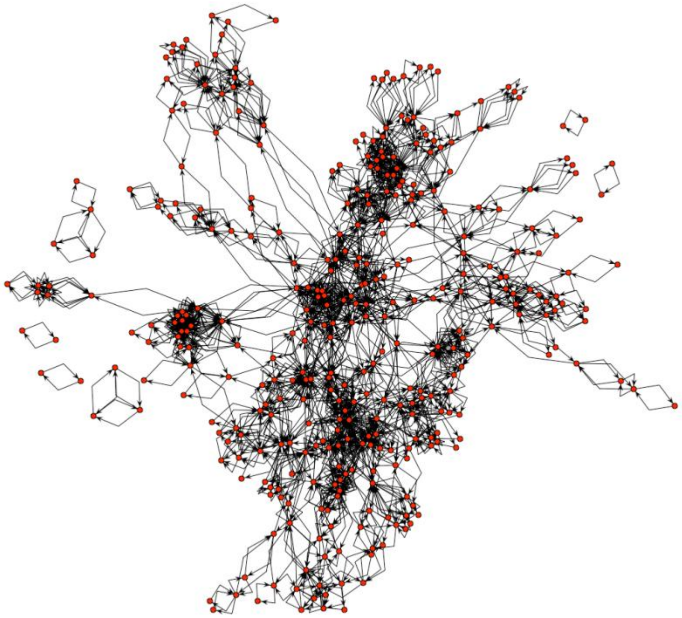
Dave Hunter
Penn State

Collaboration Network



Collaboration Network





Data

↓

Models

↓

Predictions

Statistical Modeling of Network Data

Statistics = principled approach for inference from noisy data

Basis for optimal prediction
- computation of conditional probabilities/expectation

Principles for handling noisy measurements
- e.g., noisy and missing edges

Integration of different sources of information
- e.g., combining edge information with node covariates

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

Limitations of Existing Methods

- 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*

Example

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

  V = set of N nodes

  E = set of directed binary edges

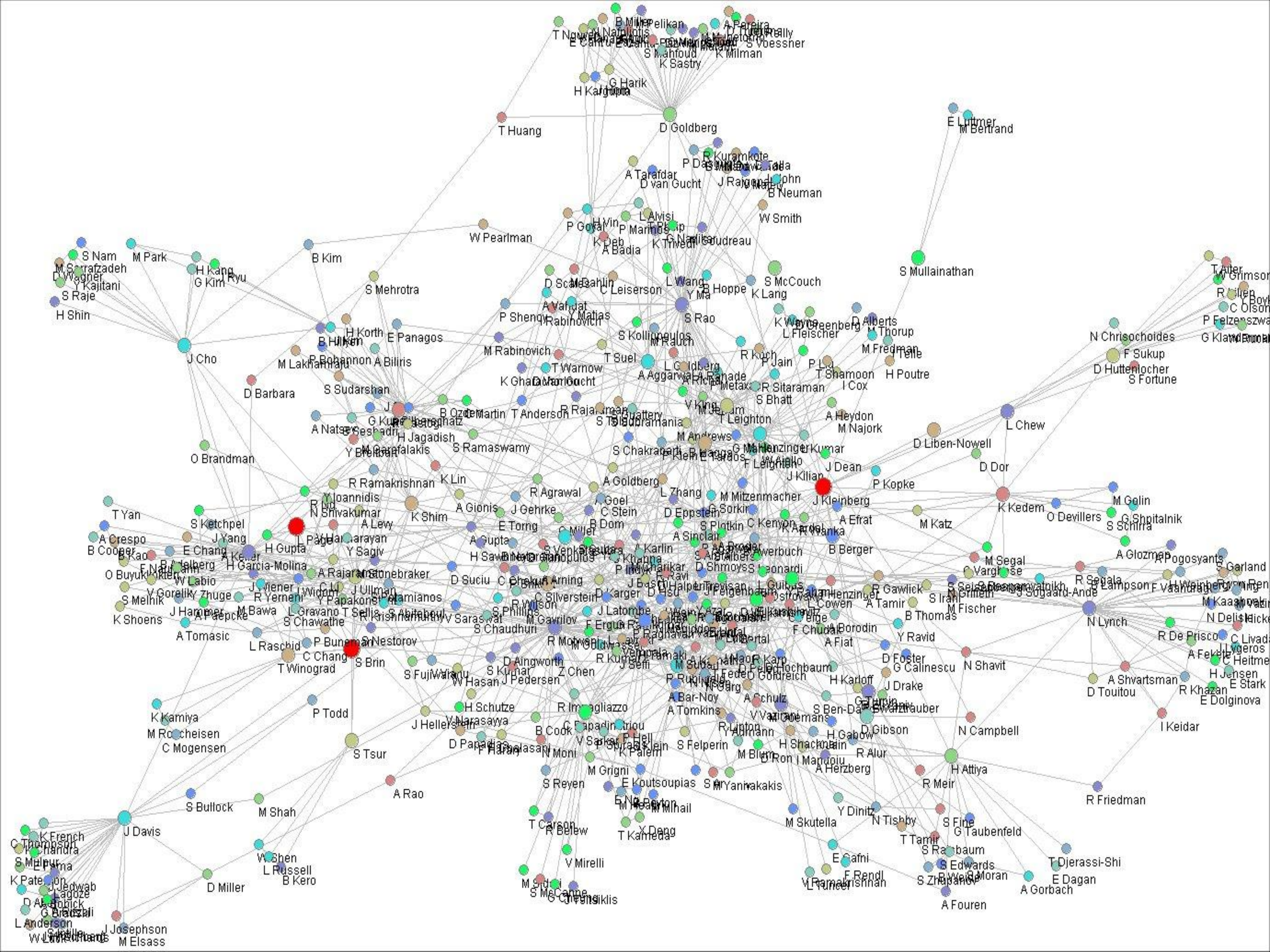
- Exponential random graph (ERG) model

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

The normalization constant = sum over all possible graphs

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

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

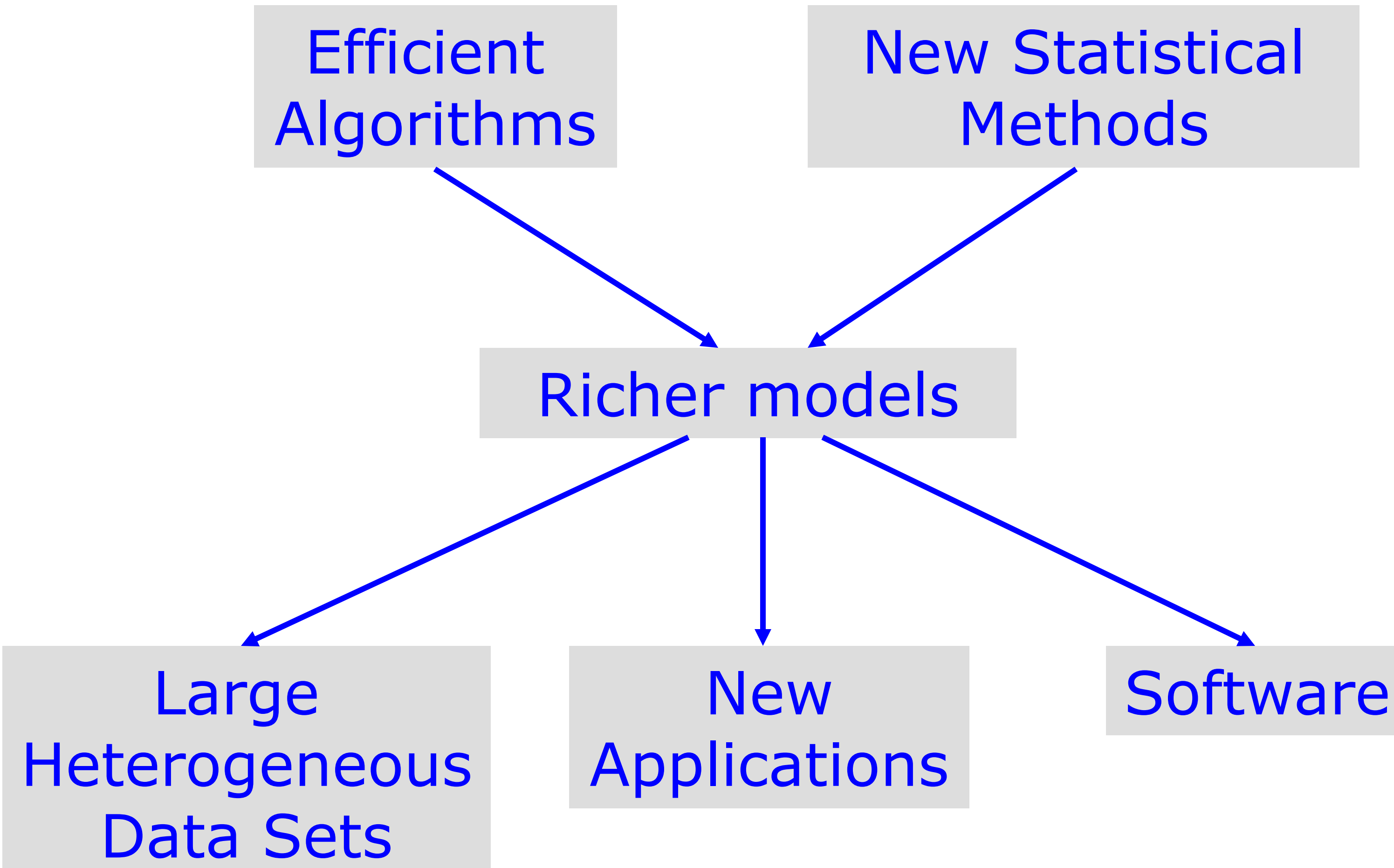


D Goldberg
T Huang
K Sastry
S Mahfoud
K Milman
S Voessner
E Luttmer
M Bertrand
S Mullainathan
W Pearlman
S Nam
M Park
M Sarrafzadeh
D Wagner
Y Kajitani
S Raje
H Shin
H Kang
G Kim
B Kim
J Cho
S Mehrotra
H Korth
E Panagos
M Lakhamraju
A Biliris
D Barbara
S Sudarshan
O Brandman
A Tarafdar
D van Gucht
B Neuman
W Smith
P Goyal
P Marinos
A Badia
K Deb
C Leiserson
S McCouch
K Lang
S Rao
P Shenoy
M Rabinovich
T Warnow
T Suel
L Fleischer
M Fredman
H Poutre
I Cox
T Shamoon
T Leighton
R Sitaraman
S Bhatt
A Heydon
M Najork
D Liben-Nowell
L Chew
D Dor
J Dean
P Kopke
J Kleinberg
K Kedem
O Devillers
S Schirra
M Katz
A Efrat
B Berger
N Chrisochoides
F Sukup
D Huttenlocher
S Fortune
R Ramakrishnan
K Lin
K Shim
A Gionis
R Agrawal
J Gehrke
Y Ioannidis
N Shivakumar
A Lew
L Page
S Brin
T Winograd
C Chang
P Buneman
S Nestorov
S Tsur
A Rao
T Yan
A Crespo
B Cooper
S Ketchpel
H Gupta
H Garcia-Molina
J Ullman
A Tomasic
L Raschid
S Chawathe
S Melnik
K Shoens
J Hammer
M Bawa
Y Sagiv
D Suciu
S Chaudhuri
M Gavrilov
R Motwani
P Indyk
D Karger
J Latombe
D Shmoys
E Tardos
L Guibas
M Segal
S Irani
M Fischer
N Lynch
N Shavit
D Touitou
A Shvartsman
I Keidar
E Dolginova
R Friedman
H Attiya
N Campbell
G Calinescu
H Karloff
J Drake
D Gibson
A Tomkins
A Bar-Noy
R Alur
M Blum
A Herzberg
R Meir
S Fine
G Taubenfeld
T Djerassi-Shi
E Dagan
A Gorbach
A Fouren
S Edwards
S Rajsbaum
T Tamir
N Tishby
Y Dinitz
M Skutella
E Gafni
F Rendl
M Yannakakis
E Koutsoupias
S Reyen
M Grigni
T Carson
R Belew
V Mirelli
X Deng
T Kameda
J Hellerstein
V Narasayya
H Schutze
J Pedersen
W Hasan
Z Chen
D Aingworth
R Impagliazzo
S Felperin
K Kamiya
M Rotcheisen
C Mogensen
S Bullock
M Shah
J Davis
W Shen
L Russell
B Kero
D Miller
K French
J Josephson
M Elsass
L Anderson

Key Themes of our MURI Project

- Foundational research on new statistical estimation techniques for network data
  - e.g., principles of modeling with missing data
- Faster 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)

Key Themes of our MURI Project



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

Task A: Fast Network Estimation Algorithms

Investigators: Eppstein, Butts

- Problem:
  - Statistical inference algorithms can be slow because of repeated computation of various statistics on graphs
- Goal
  - Leverage ideas from computational graph algorithms to enable much faster computation – also enabling computation of more complex and realistic statistics
- Projects
  - Dynamic graph methods for change-score computation
  - Rapid subgraph automorphism detection for feature counting
  - Dynamic connectivity

Task B: Spatial Representations and Network Data

Investigators: Goodrich, Eppstein, Mount

- Problem:
  - Spatial representations of network data can be quite useful (both latent embeddings and actual spatial information) but current statistical modeling algorithms scale poorly
- Goal
  - Build on recent efficient geometric data indexing techniques in computer science to develop much faster and efficient algorithms
- Projects
  - Improved algorithms for latent-space embeddings
  - Fast implementations for high-dimensional latent space models
  - Techniques for integrating actual and latent space geometry

Task C: Advanced Estimation Techniques

Investigators: Handcock, Hunter

- Problem:
  - Current statistical network inference models often make unrealistic assumptions, e.g.,
    - Assume complete (non-missing) data
    - Assume that exact computation is possible
- Goal
  - Develop new theories and techniques that relax these assumptions, i.e., methods for handing missing data and techniques for approximate inference
- Projects
  - Inference with partially observed network data
  - Approximation methods
    - Approximate likelihood techniques
    - Approximate MCMC algorithms
  - Will leverage new techniques developed in Tasks A and B

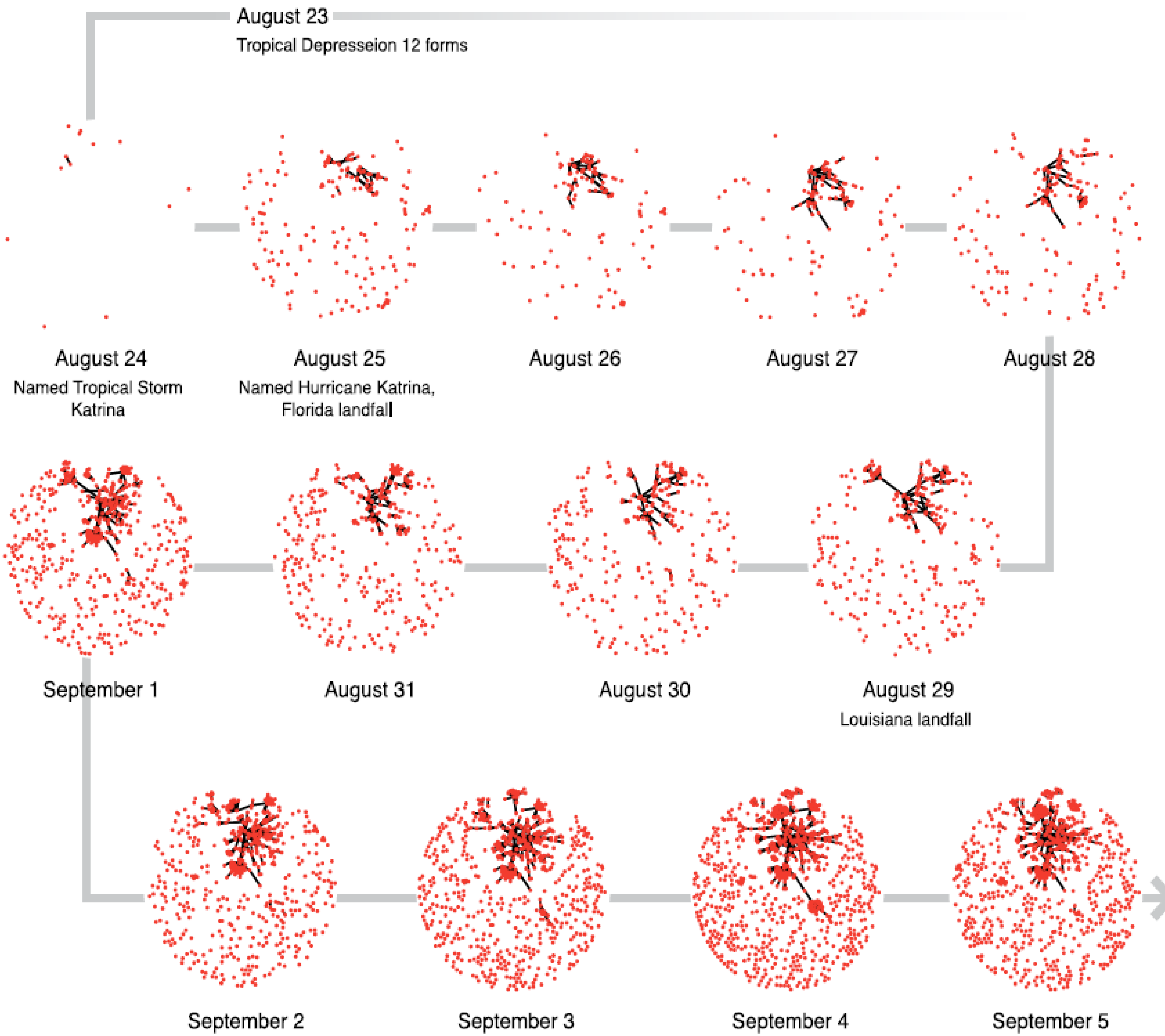
Task D: Scalable Temporal Models

Investigator: Butts

- Problem:
  - Few statistical methods for modeling temporal sequences of events among a network of actors

- Goal
  - Develop new statistical relational event models to handle an evolving set of events over time in a network context

- Projects
  - Specification of relational event statistics
  - Rapid likelihood computation for relational event models
  - Predictive event system queries
  - Interventions, forecasting, and "network steering"
  - Can build on ideas from Tasks A, B, C

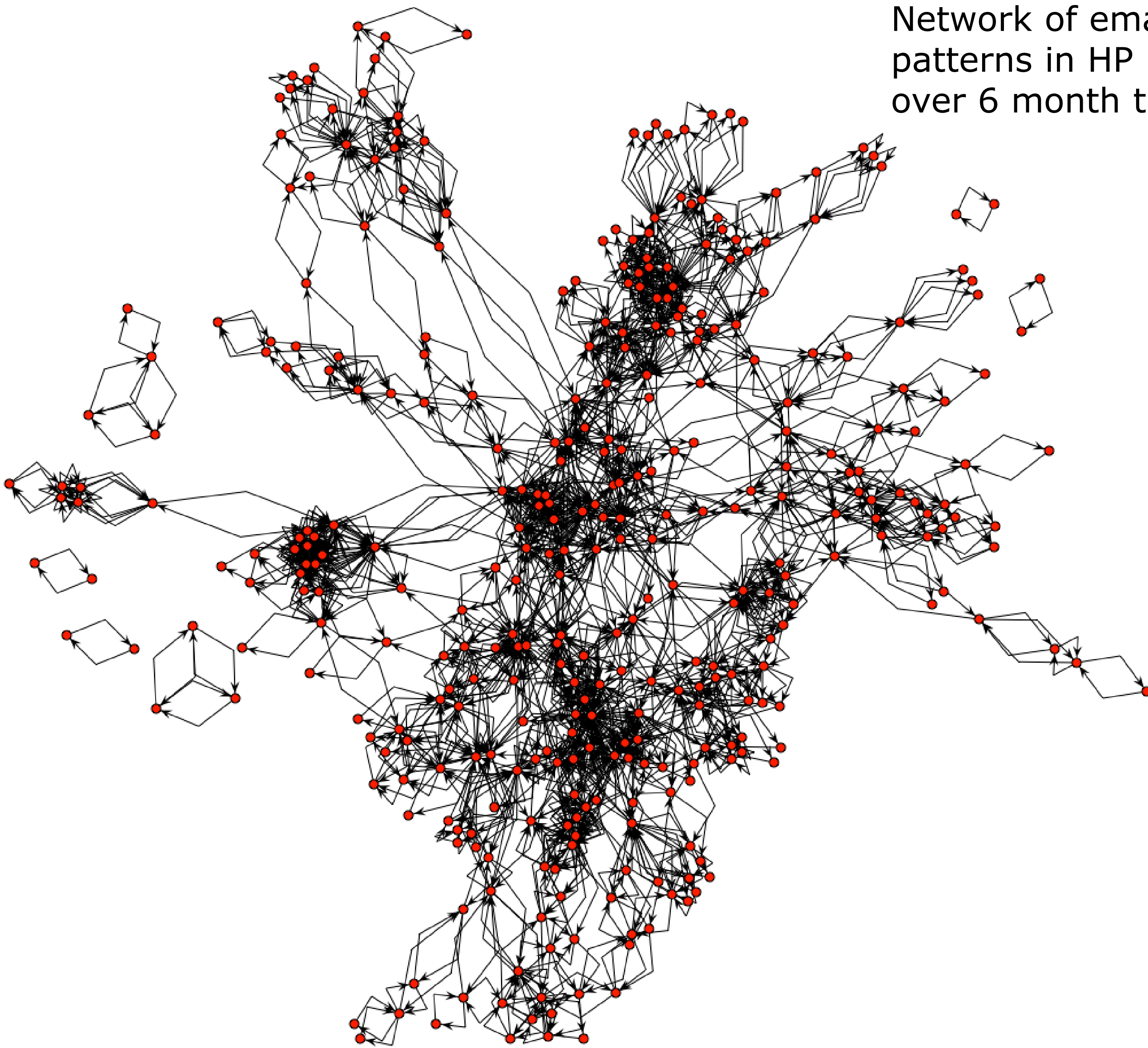

August 23
Tropical Depresseion 12 forms
August 24
Named Tropical Storm Katrina
August 25
Named Hurricane Katrina, Florida landfall
August 26
August 27
August 28
September 1
August 31
August 30
August 29
Louisiana landfall
September 2
September 3
September 4
September 5

Task E: Network Models and Text Data

Investigator: Smyth

- Problem:
  - Lack of statistical techniques that can combine network and text data within a single framework (e.g., email communication)
- Goal
  - Leverage recent advances in both statistical text mining and statistical network modeling to create new combined models
- Projects
  - Latent variable models for text and network data
  - Text as exogenous data for statistical network models
  - Modeling of text and network data over time
  - Fast algorithms for statistical modeling of text/networks
  - Can build on ideas from Tasks A, B, C and D

Network of email communication patterns in HP Research Labs over 6 month time-frame



Task F: Software for Network Inference and Prediction

Investigator: Hunter

- Goal
  - Disseminate algorithms and software to research and practitioner communities

- How?
  - By incorporating our new algorithms into the R statistical package
  - R = open source language for stat computing/graphics
  - MURI team has significant prior experience with developing statistical network modeling packages in R
    - *network* (Butts et al, 2007)
    - *latentnet* (Handcock et al, 2004)
    - *ergm* (Handcock et al, 2003)
    - *sna* (Butts, 2000)

- Will integrate algorithms and techniques from other tasks

ONR Interests

(adapted from presentation/discussion 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?

Timelines and Funding

- 3-year project, possible extension to 5 years
  - Start date: May 1 2008
  - End date: April 30 2011/2013
- Funding installment 1:
  - First 5 months of funding, intended for May-Sept 2008
  - Arrived at UCI in Sept 2008
  - Largely spent by March 2008
- Funding installment 2:
  - 12 months of funding, intended for Oct 1 08 to Sep 30 09
  - Arrived at UCI mid-march 2009
  - Plan to spend current funding by March 2010
- Anticipate next installment will arrive in early 2010

Project Meetings

- All-Hands Meeting, November 2008
  - Researchers + ONR program manager (Martin Kruger) + other DoD folks
- Working Meeting, April 2009
  - Researchers
- Working Meeting, August 2009
  - Researchers + Julie Howell and Joan Kaina (Navy, San Diego)
- All-Hands Meeting, November 2009
  - Researchers + program manager + other DoD folks
  - Exact date TBD

Research Examples

- Statistical modeling of network data with missing observations
  - Mark Handcock and Krista Gile
  - Systematic statistical methodologies for handling missing edge information in observed network data
- Decision-theoretic foundations for network modeling
  - Carter Butts
  - Network formation via stochastic choice processes and links to exponential random graph (ERG) models
- Fast computation of graph change scores in large networks
  - David Eppstein and Emma Spiro
  - New data structure that significantly speeds up the evaluation of change-score statistics in ERG estimation

Sample 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

Sample Publications (ctd.)

- M. Gjoka, M. Kurant, C. T. Butts, A. Markopoulou, A walk in Facebook: uniform sampling of users in online social networks, electronic preprint, arXiv:0906.0060, 2009

- M. Cho, D. M. Mount, and E. Park, Maintaining nets and net trees under incremental motion, submitted, 2009

- 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

Morning Session I

9:30 Foundational aspects of network analysis
Carter Butts (UCI)

9:45 Comparison of estimation methods for exponential random graph models
Mark Handcock (UW)

10:15 Sampling algorithms for data collection in online networks
Carter Butts (UCI)

10:30 Break

Morning Session II

10:45 Egocentric network models for event data over time
Chris Marcum, Lorien Jasny, Carter Butts (UCI)

11:15 Dynamic extensions of network brokerage models
Ryan Acton, Emma Spiro, Carter Butts (UCI)

11:30 Statistical approaches to joint modeling of text and network data
Arthur Asuncion, Qiang Liu, Padhraic Smyth (UCI)

12:00 Lunch for all at University Club

Afternoon Session I

1:30 The crossroads of geography and networks
Michael Goodrich (UCI)

2:00 Maintaining nets and net trees under incremental motion
Minkyoung Cho, Eunhui Park, Dave Mount (U Maryland)

2:30 Simulation of spatially-embedded network data
Carter Butts (UCI)

3:00 A proposal for the analysis of disaster-related network data, Miruna Petrescu-Prahova (UW)

3:30 Break

Afternoon Session II

3:45 Approximate inference techniques with applications to spatial network models

Drew Frank, Alex Ihler, Padhraic Smyth (UCI)

4:15 Update on project data organization, assembly, and collection

Emma Spiro (UCI)

4:30 Discussion and Wrap-up
- date of AHM meeting in November
- collaborative activities
- action items

5:00 Adjourn

Logistics

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