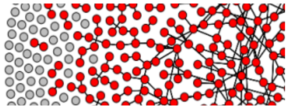


Scalable Methods for the Analysis of Network-Based Data

MURI Meeting, June 3rd 2011, UC Irvine

Principal Investigator:
Professor Padhraic Smyth
Department of Computer Science
University of California, Irvine

Additional project information online at www.datalab.uci.edu/muri

Today's Meeting

- Goals
  - Review our research progress
  - Discussion, questions, interaction
  - Feedback from visitors
- Format
  - Introduction
  - Research talks
    - Regular: 20 minutes + 5 mins at end for questions/discussion
    - Short: 10 minutes (session after lunch)
  - Two open discussion sessions, led by faculty
  - Question/discussion encouraged during talks
  - Several breaks for discussion

MURI Project Timeline

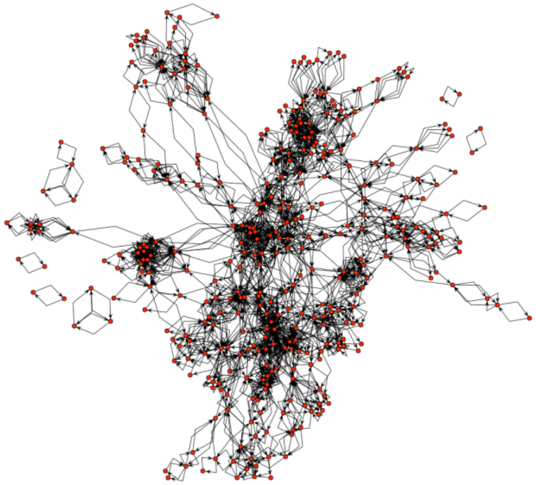
- Initial 3-year period
  - May 1 2008 to April 30th 2011
  - Funding actually arrived to universities in Oct 2008

- 2-year extension:
  - May 1 2011 to April 30th 2013

- Meetings (all at UC Irvine)
  - Kickoff Meeting, November 2008
  - Working Meetings, April 2009, August 2009
  - Annual Review, December 2009
  - Working Meeting, May 2010
  - Annual Review, November 2010
  - Today, June 2011

Motivation

2007: interdisciplinary interest in analysis of large network data sets

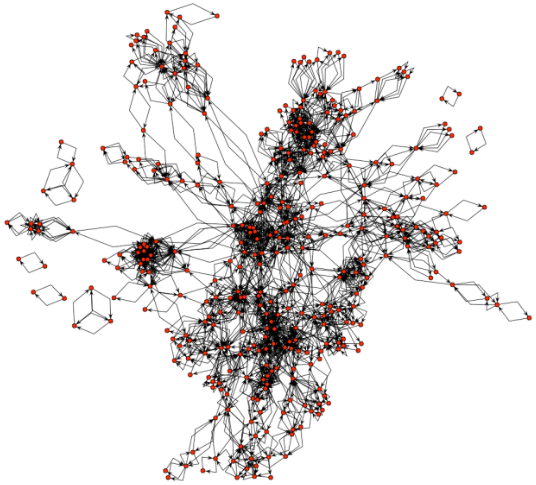


Many of the available techniques were descriptive, could not handle

- Prediction
- Missing data
- Covariates, etc

Motivation

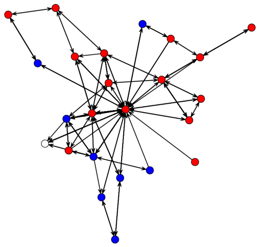
2007: interdisciplinary interest in analysis of large network data sets



Many of the available techniques were descriptive, could not handle

- Prediction
- Missing data
- Covariates, etc

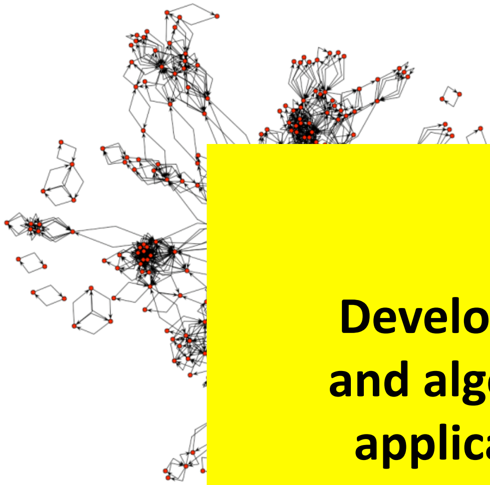
2007: significant statistical body of theory available on network modeling



Many of the available techniques did not scale up to large data sets, not widely known/understood/used, etc

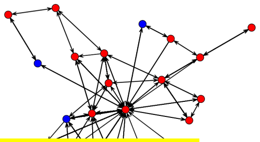
Motivation

2007: interdisciplinary interest in analysis of large network data sets



Many of the avai
descriptive,
- Prediction
- Missing data
- Covariates, etc

2007: significant statistical body of theory available on network modeling



chniques did not
a sets, not widely
used, etc

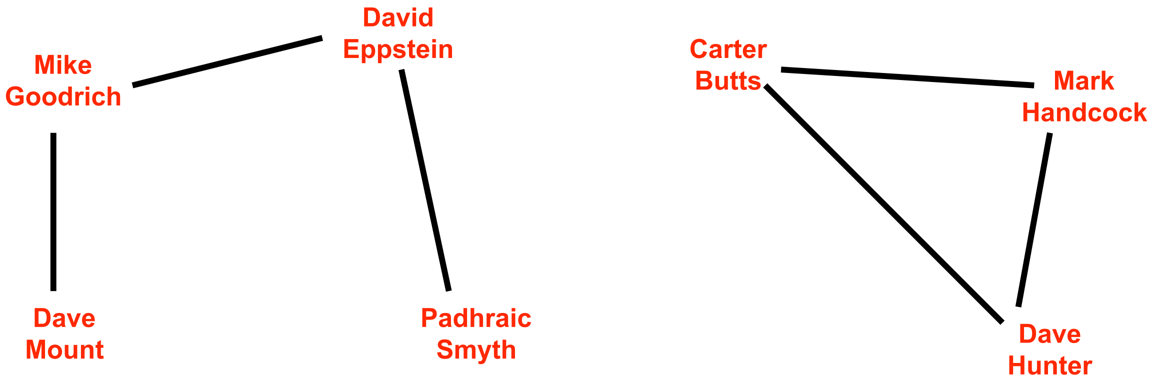
Goal of this MURI project

Develop new statistical network models and algorithms to broaden their scope of application to large, complex, dynamic real-world network data sets

MURI Team

Investigator	University	Department	Expertise	Number Of PhD Students	Number of Postdocs
Padhraic Smyth (PI)	UC Irvine	Computer Science	Machine learning	4	
Carter Butts	UC Irvine	Sociology	Statistical social network analysis	6	
Mark Handcock	UCLA	Statistics	Statistical social network analysis	1	1
Dave Hunter	Penn State	Statistics	Computational statistics	2	1
David Eppstein	UC Irvine	Computer Science	Graph algorithms	2	
Michael Goodrich	UC Irvine	Computer Science	Algorithms and data structures	1	1
Dave Mount	U Maryland	Computer Science	Algorithms and data structures	2	
			TOTALS	**18**	**3**

Collaboration Network



David Eppstein

Mike Goodrich

Dave Mount

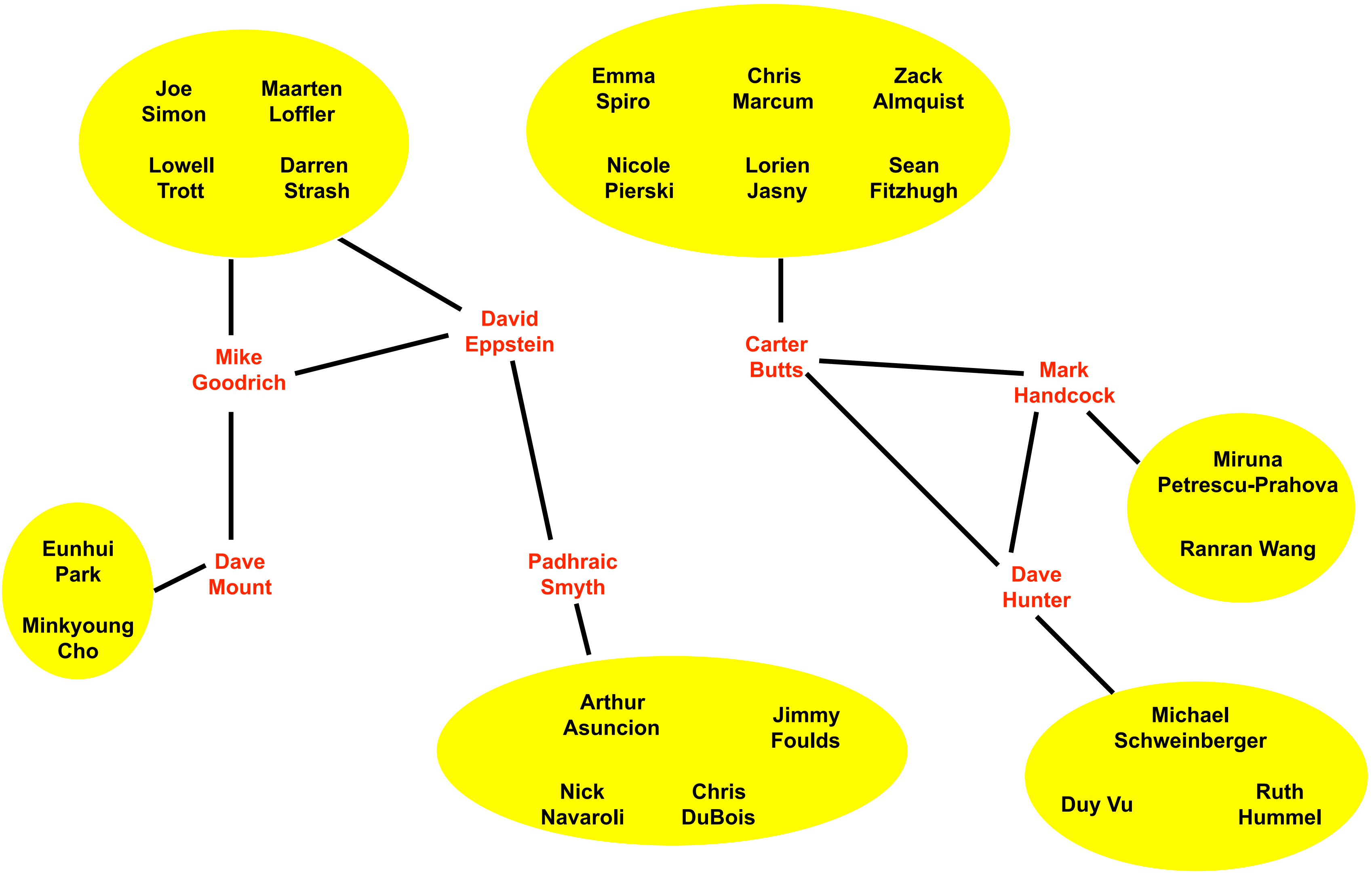
Padhraic Smyth

Carter Butts

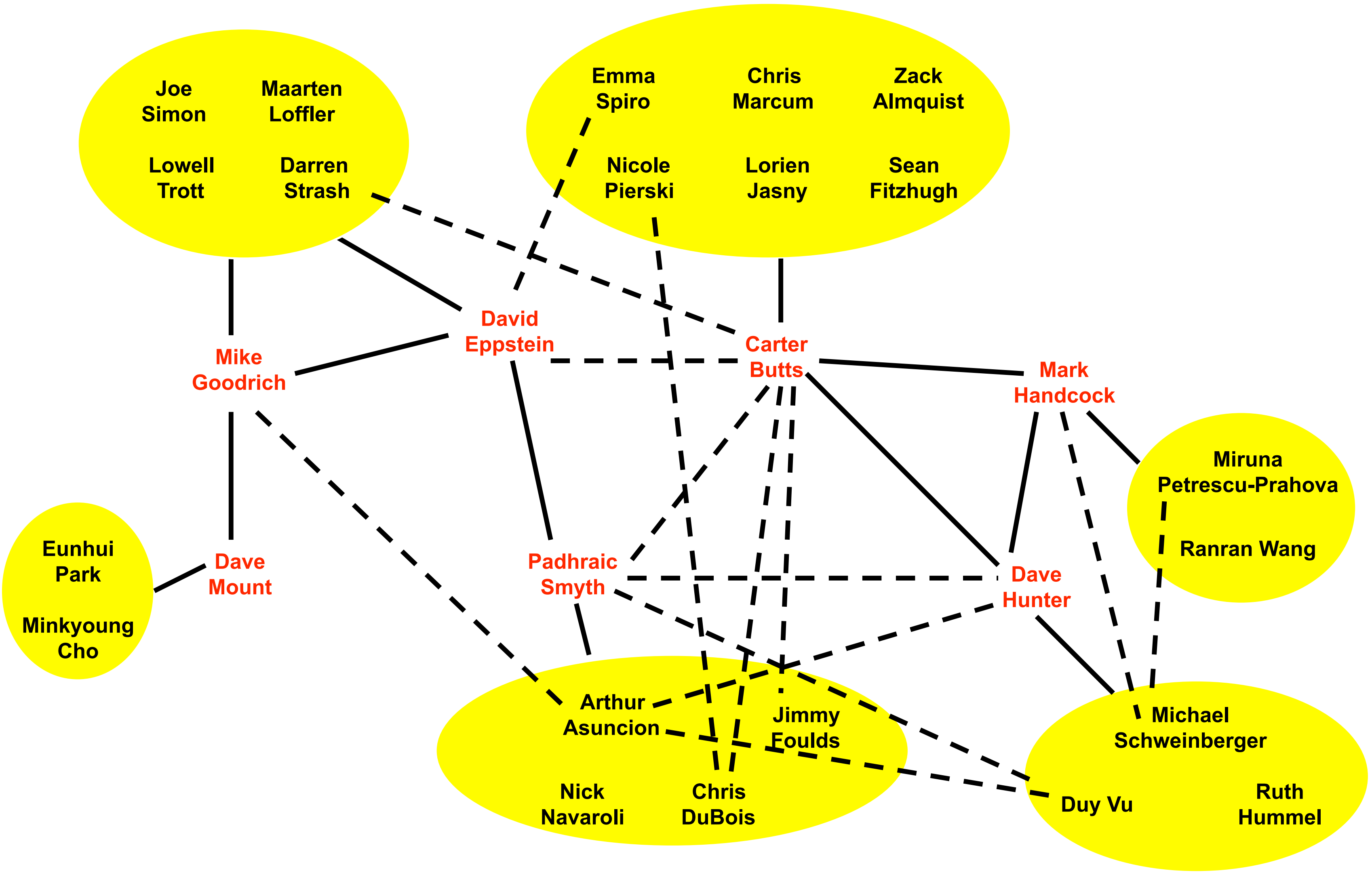
Mark Handcock

Dave Hunter

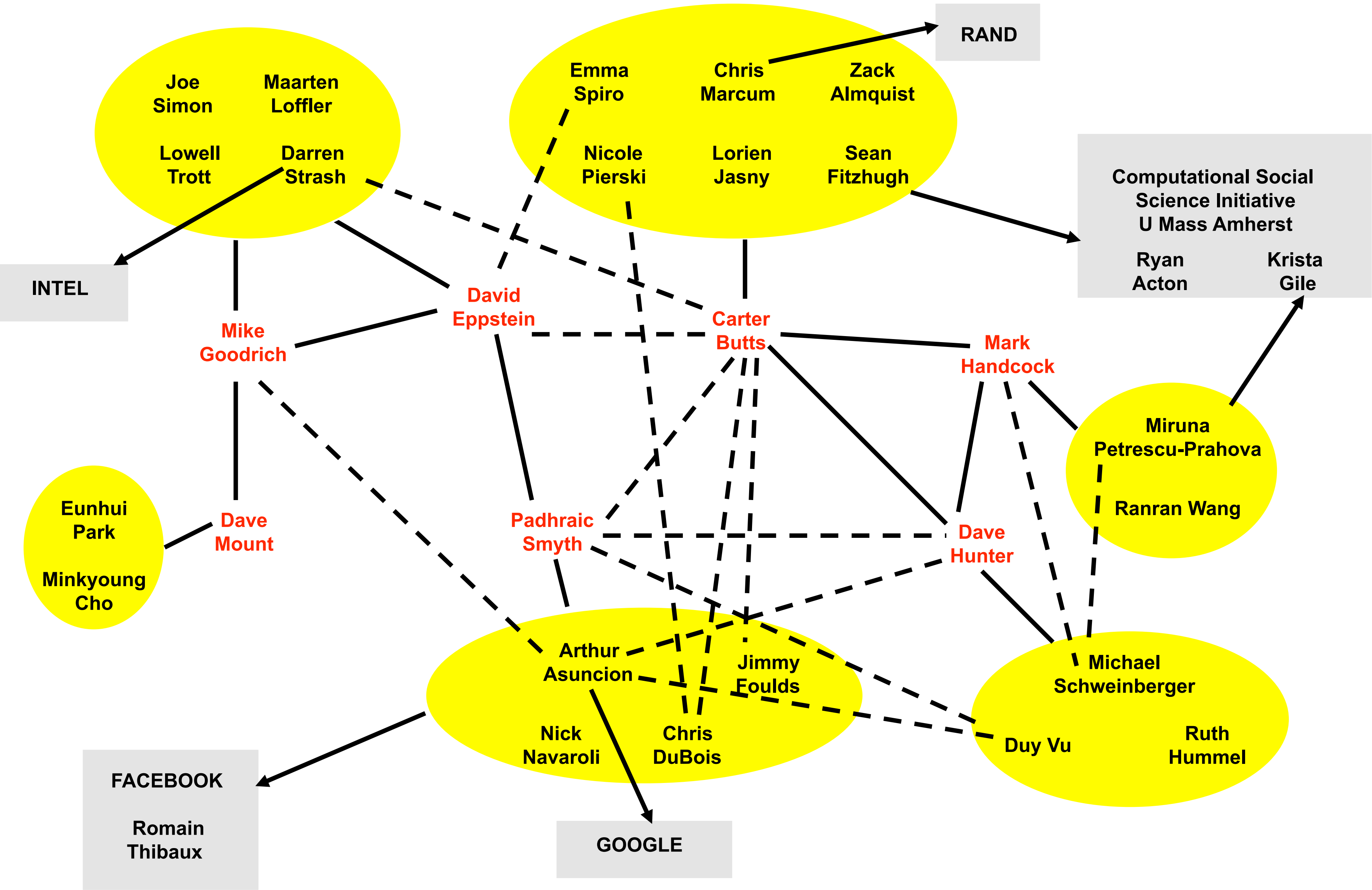
Collaboration Network



Collaboration Network



Collaboration Network



Joe Simon

Maarten Loffler

Lowell Trott

Darren Strash

Emma Spiro

Chris Marcum

Zack Almquist

Nicole Pierski

Lorien Jasny

Sean Fitzhugh

RAND

INTEL

Computational Social Science Initiative U Mass Amherst

Ryan Acton

Krista Gile

David Eppstein

Mike Goodrich

Carter Butts

Mark Handcock

Miruna Petrescu-Prahova

Ranran Wang

Eunhui Park

Minkyoung Cho

Dave Mount

Padhraic Smyth

Dave Hunter

Arthur Asuncion

Jimmy Foulds

Nick Navaroli

Chris DuBois

Michael Schweinberger

Duy Vu

Ruth Hummel

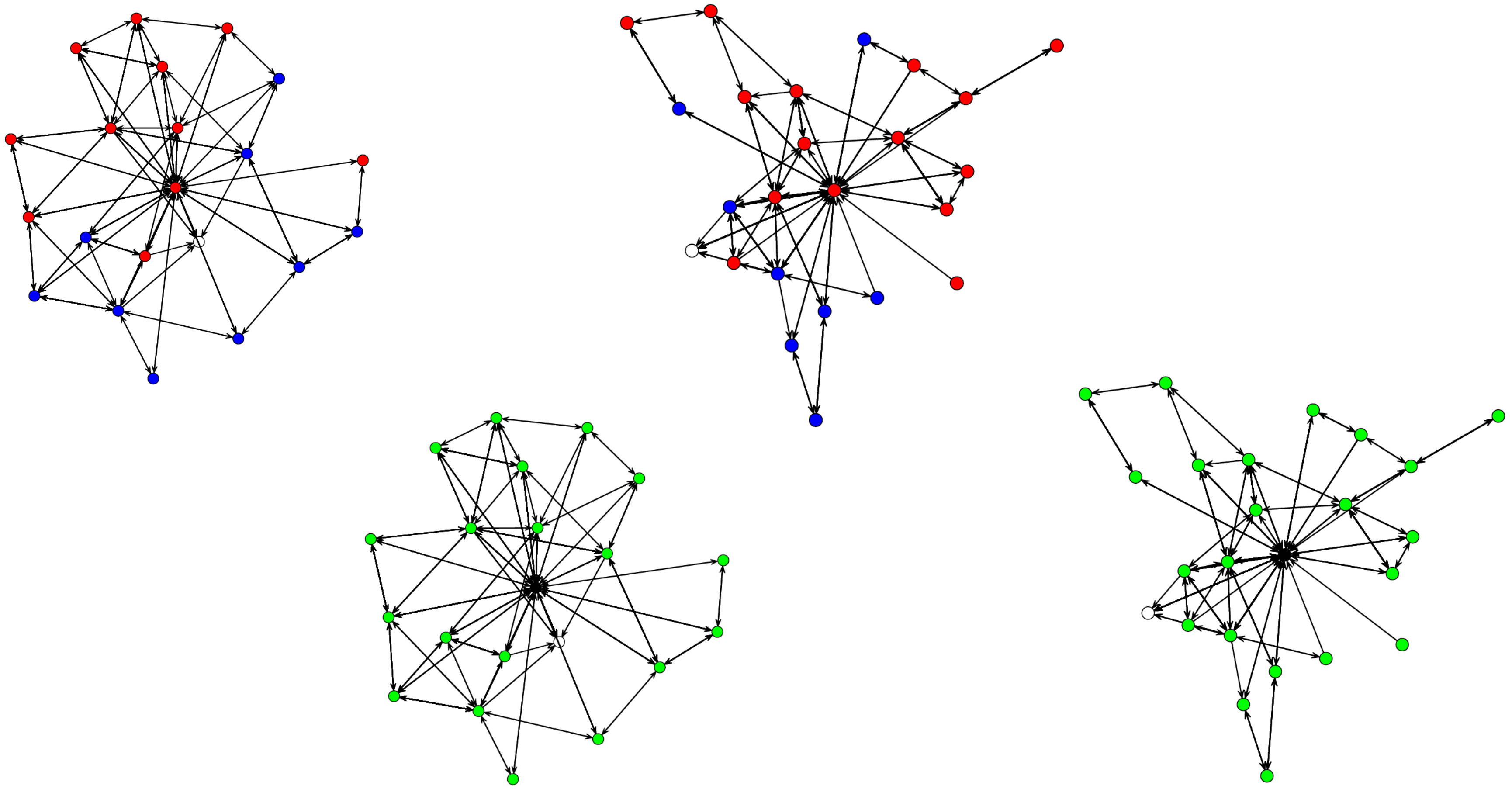
FACEBOOK

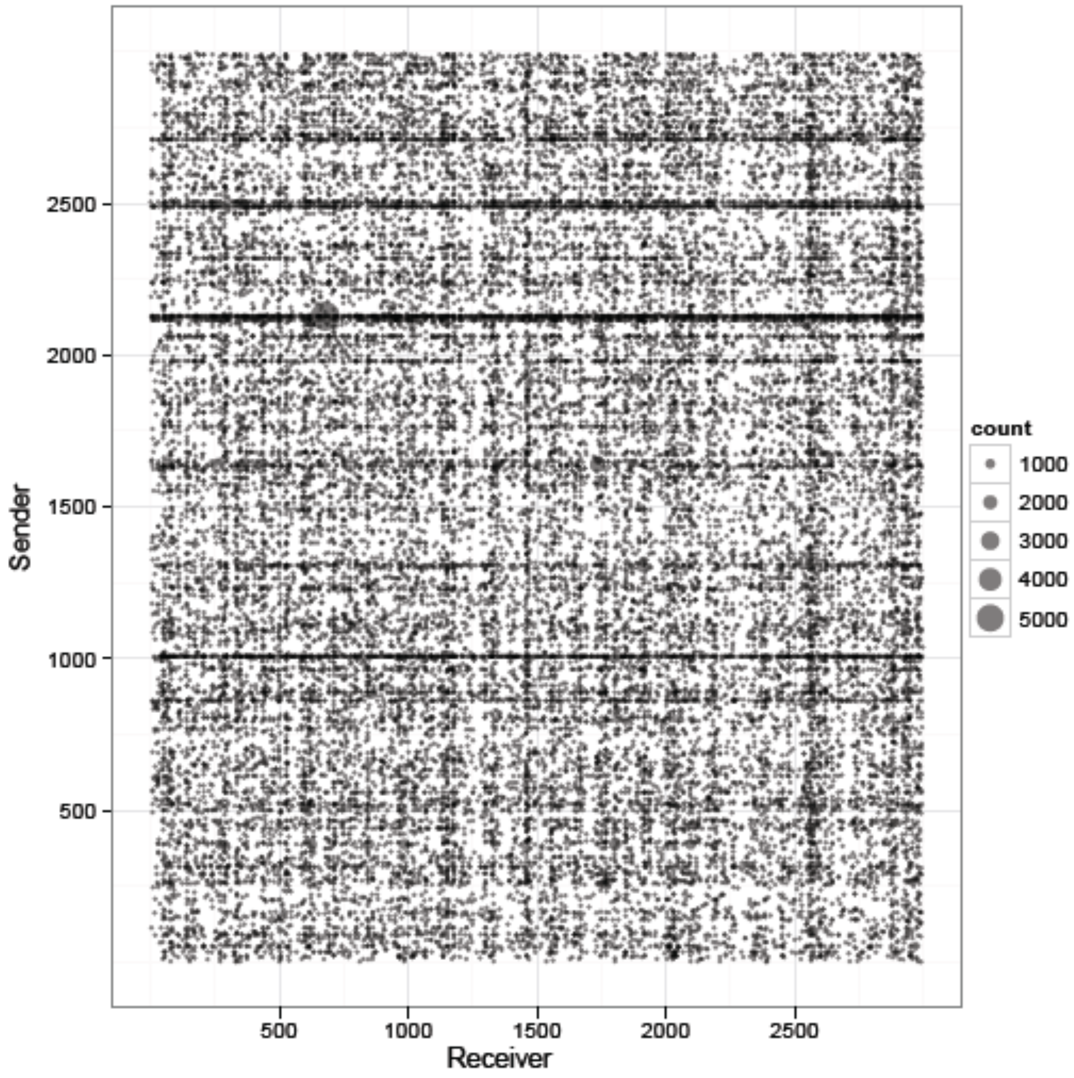
Romain Thibaux

GOOGLE

Example: Network Dynamics in Classrooms

Nicole Pierski, Chris DuBois, Carter Butts

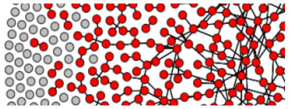




Data:
Count matrix of 200,000 email messages among 3000 individuals over 3 months

Problem:
Understand communication patterns and predict
future communication activity

Challenges:
sparse data, missing data,
non-stationarity,
unseen covariates

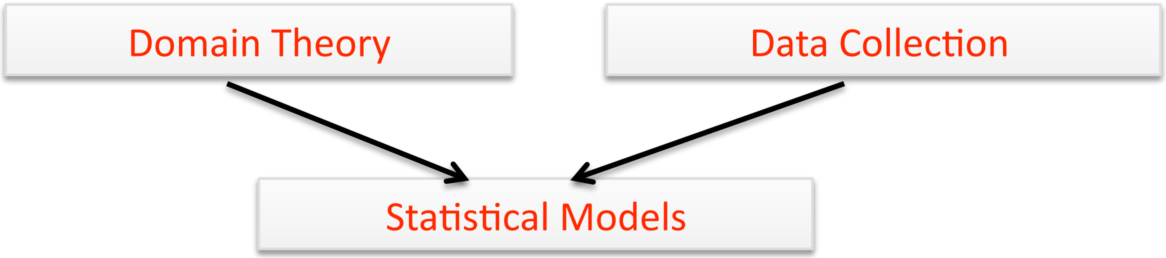


Key Scientific/Technical Challenges

- Parametrize models in a sensible and computable way
  - Respect theories of social behavior as well as explain observed data
- Account for real data
  - E.g., understand sampling methods: account for missing, error-prone data
- Make inference both principled and practical
  - computationally-scalability: want accurate conclusions, but can't wait forever for results
- Deal with rich and dynamic data
  - Real-world problems involve systems with complex covariates (text, geography, etc) that change over time

In sum: statistically principled methods that respect the realities of data and computational constraints

Mapping the Project Terrain

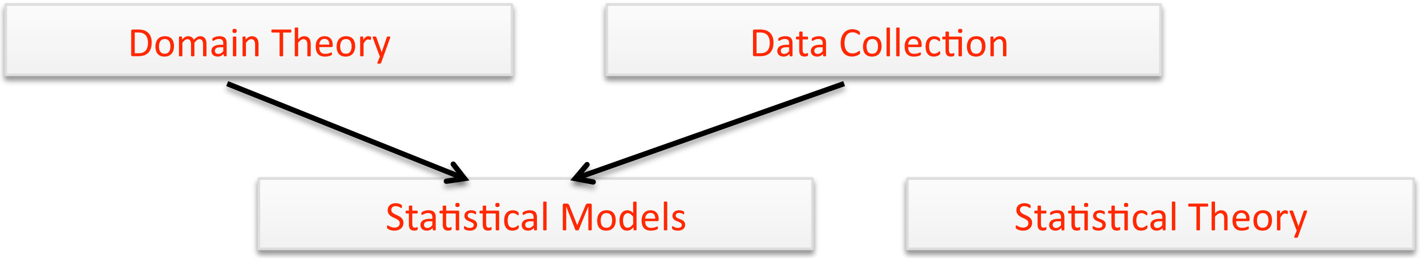


Domain Theory

Data Collection

Statistical Models

Mapping the Project Terrain



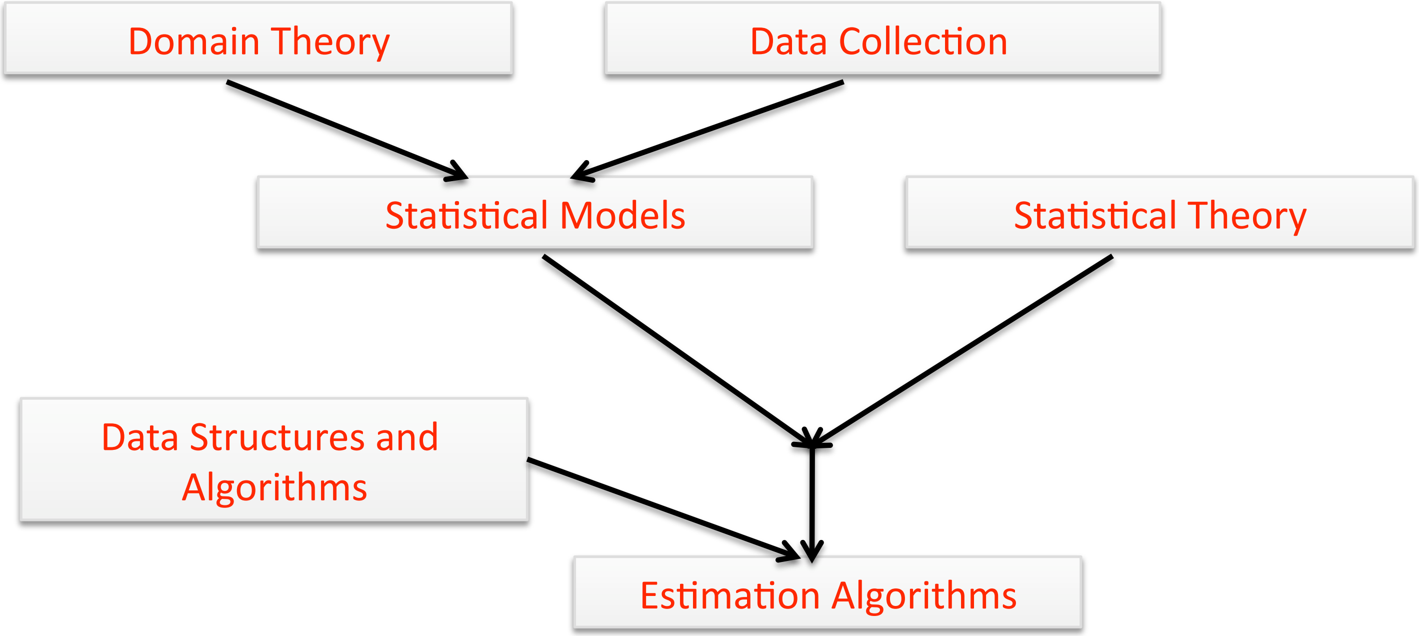
Domain Theory

Data Collection

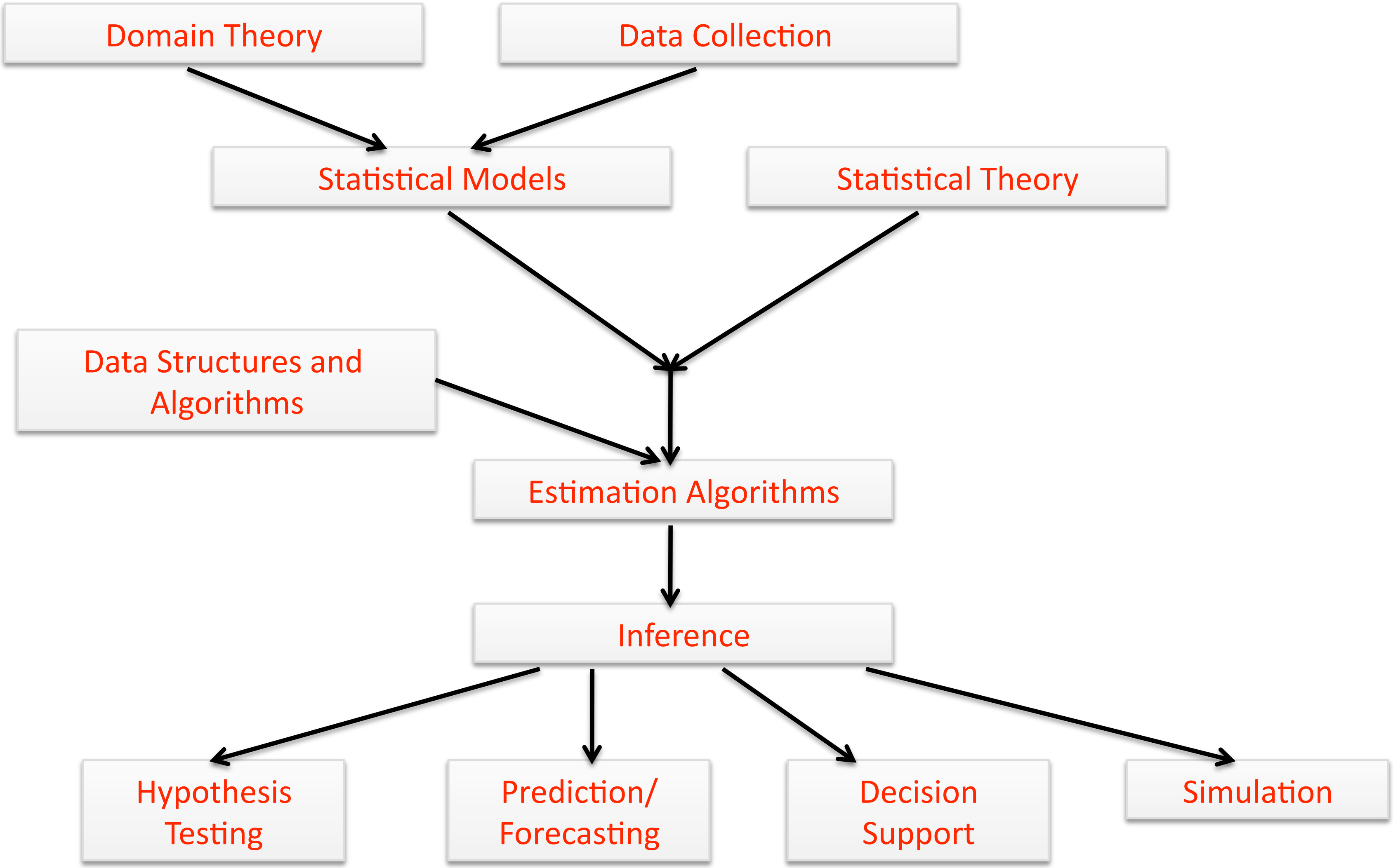
Statistical Models

Statistical Theory

Mapping the Project Terrain



Mapping the Project Terrain



Domain Theory

Data Collection

Statistical Models

Statistical Theory

Data Structures and Algorithms

Estimation Algorithms

Inference

Hypothesis Testing

Prediction/ Forecasting

Decision Support

Simulation

Accomplishments

Topic	State of the Art Prior to Project	State of the Art Now	Potential Applications And Impact
General theory for handling missing data in social networks	Problem only partially understood. No software available for statistical modeling	General statistical theory for treating missing data in a social network context. Publicly-available code in R. (Gile and Handcock, 2010)	Allows application of social network modeling to data sets with significant missing data
Relational event models	Basic dyadic event models. No exogenous events. No public software.	Much richer model with exogenous events, egocentric support, multiple observer accounts, hierarchies Software publicly available (Butts et al, 2010)	Provides a general framework for dynamic network modeling to large realistic applications
Clique finding algorithms	Too slow for use in statistical network modeling	New linear-time algorithm for listing all maximal cliques in sparse graphs (Eppstein, Loffler, Strash, 2010)	Extends applicability of statistical network modeling to larger networks and more complex models

Accomplishments

Topic	State of the Art Prior to Project	State of the Art Now	Potential Applications And Impact
General theory for handling missing data in social networks	Problem only partially understood. No software available for statistical modeling	General statistical theory for treating missing data in a social network context. Publicly-available code in R. (Gile and Handcock, 2010)	Allows application of social network modeling to data sets with significant missing data
Relational event models	Basic dyadic event models. No exogenous events. No public software.	Much richer model with exogenous events, egocentric support, multiple observer accounts, hierarchies Software publicly available (Butts et al, 2010)	Provides a general framework for dynamic network modeling to large realistic applications
Clique finding algorithms	Too slow for use in statistical network modeling	New linear-time algorithm for listing all maximal cliques in sparse graphs (Eppstein, Loffler, Strash, 2010)	Extends applicability of statistical network modeling to larger networks and more complex models

Accomplishments

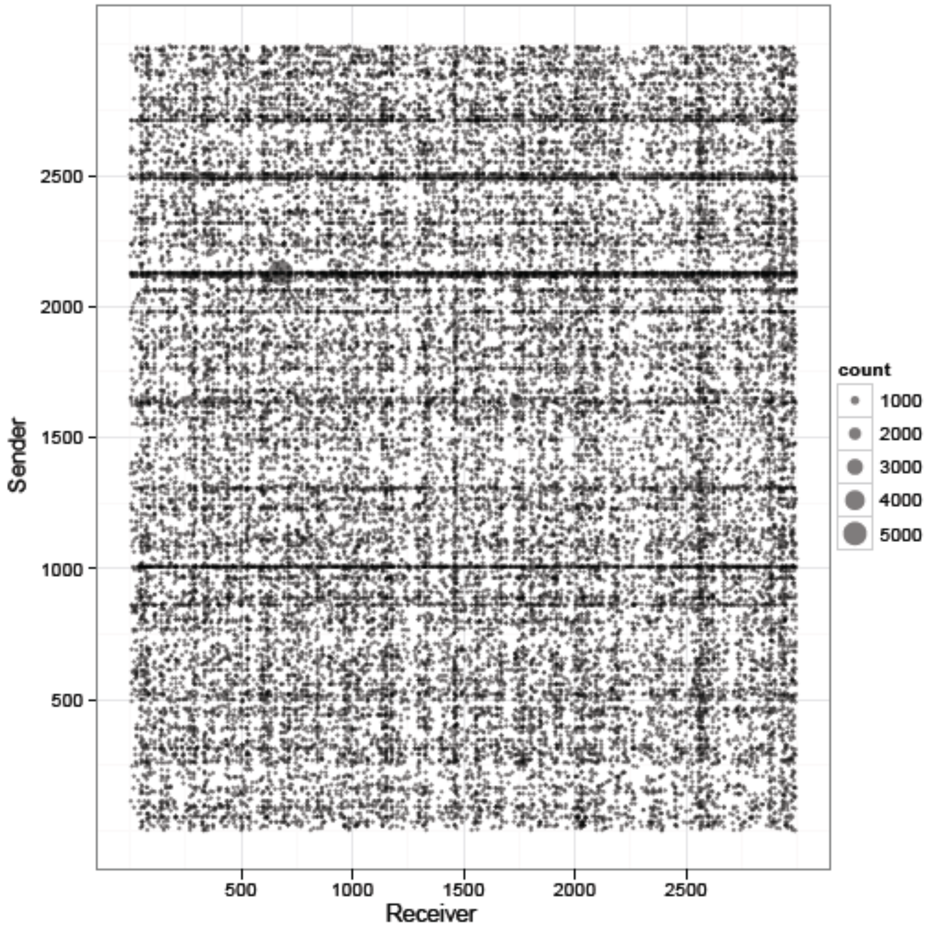
Topic	State of the Art Prior to Project	State of the Art Now	Potential Applications And Impact
General theory for handling missing data in social networks	Problem only partially understood. No software available for statistical modeling	General statistical theory for treating missing data in a social network context. Publicly-available code in R. (Gile and Handcock, 2010)	Allows application of social network modeling to data sets with significant missing data
Relational event models	Basic dyadic event models. No exogenous events. No public software.	Much richer model with exogenous events, egocentric support, multiple observer accounts, hierarchies Software publicly available (Butts et al, 2010)	Provides a general framework for dynamic network modeling to large realistic applications
Clique finding algorithms	Too slow for use in statistical network modeling	New linear-time algorithm for listing all maximal cliques in sparse graphs (Eppstein, Loffler, Strash, 2010)	Extends applicability of statistical network modeling to larger networks and more complex models

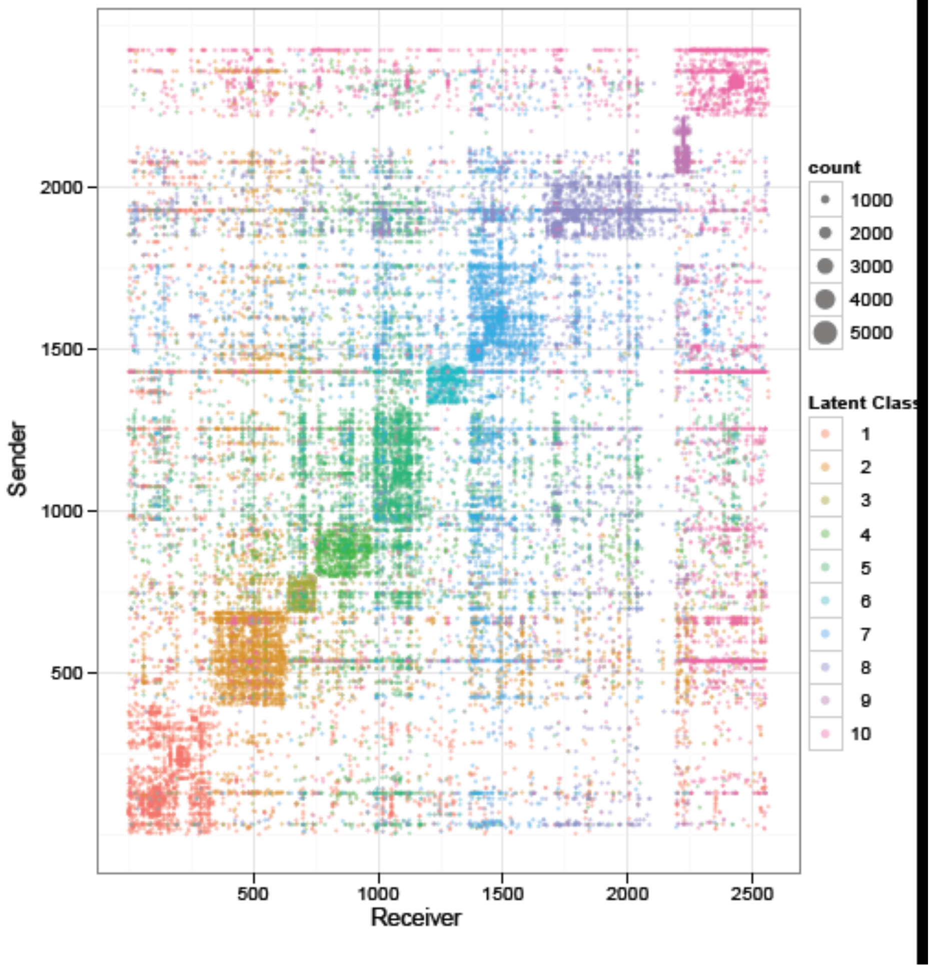
Accomplishments

Topic	State of the Art Prior to Project	State of the Art Now	Potential Applications And Impact
General theory for handling missing data in social networks	Problem only partially understood. No software available for statistical modeling	General statistical theory for treating missing data in a social network context. Publicly-available code in R. (Gile and Handcock, 2010)	Allows application of social network modeling to data sets with significant missing data
Relational event models	Basic dyadic event models. No exogenous events. No public software.	Much richer model with exogenous events, egocentric support, multiple observer accounts, hierarchies Software publicly available (Butts et al, 2010)	Provides a general framework for dynamic network modeling to large realistic applications
Clique finding algorithms	Too slow for use in statistical network modeling	New linear-time algorithm for listing all maximal cliques in sparse graphs (Eppstein, Loffler, Strash, 2010)	Extends applicability of statistical network modeling to larger networks and more complex models

Application to Email Data:

200,000 email messages among 3000 individuals over 3 months





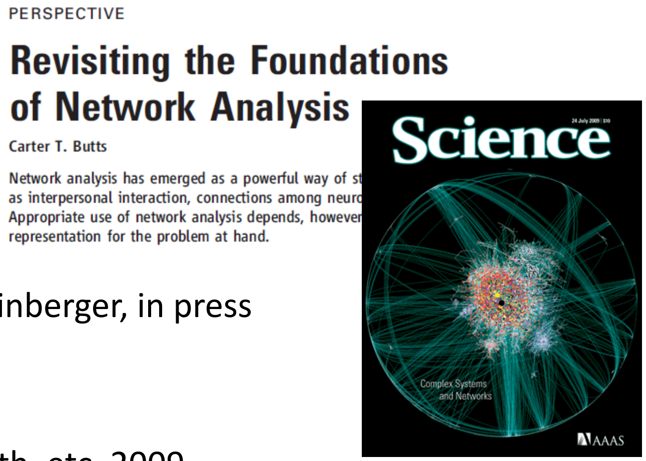
(DuBois and Smyth, ACM SIGKDD 2010)

Impact: Software

- R Language and Environment
  - Open-source, high-level environment for statistical computing
  - Default standard among research statisticians - increasingly being adopted by others
  - Estimated 250k to 1 million users

- Statnet
  - R libraries for analysis of network data
  - New contributions from this MURI project:
    - Missing data (Gile and Handcock, 2010)
    - Relational event models (Butts, 2010)
    - Latent-class models (DuBois, 2010)
    - Fast clique-finding (Strash, 2010)
    - + more……

Impact: Publications



PERSPECTIVE

Revisiting the Foundations of Network Analysis

Carter T. Butts

Network analysis has emerged as a powerful way of st
as interpersonal interaction, connections among neuro
Appropriate use of network analysis depends, however
representation for the problem at hand.

- Over 40 peer-reviewed publications
  - across computer science, statistics, and social science
  - High visibility
    - *Science,* Butts, 2009
    - *Journal of the American Statistical Association,* Schweinberger, in press
    - *Annals of Applied Statistics*, Gile and Handcock, 2010
    - *Journal of the ACM*, da Fonseca and Mount, 2010
    - *Journal of Machine Learning Research*, Asuncion, Smyth, etc, 2009
  - Highly selective conferences
    - ACM SIGKDD 2010 (16% accept rate)
    - Neural Information Processing (NIPS) Conference 2009 (25% accepts)
    - IEEE Infocom 2010 (17.5% accepts)

- Cross-pollination
  - Exposing computer scientists to statistical and social networking ideas
  - Exposing social scientists and statisticians to computational modeling ideas

Impact: Workshops and Invited Talks

- 2010 Political Networks Conference
  - Workshop on Network Analysis
  - Presented and run by Butts and students Spiro, Fitzhugh, Almquist

- Invited Talks: Conferences and Workshops
  - R!2010 Conference at NIST (Handcock, 2010)
  - 2010 Summer School on Social Networks (Butts)
  - Mining and Learning with Graphs Workshop (Smyth, 2010)
  - NSF/SFI Workshop on Statistical Methods for the Analysis of Network Data (Handcock, 2009)
  - International Workshop on Graph-Theoretic Methods in Computer Science (Eppstein, 2009)
  - Quantitative Methods in Social Science (QMSS) Seminar, Dublin (Almquist. 2010)
  - + many more…..

- Invited Talks: Universities
  - Stanford, UCLA, Georgia Tech, U Mass, Brown, etc

Impact: the Next Generation

- Faculty positions at U Mass
  - Ryan Acton, Krista Gile -> Asst Profs, part of new initiative in Computational Social Science
- Students speaking at major summer conferences
  - Sunbelt International Social Networks (Jasny, Spiro, Fitzhugh, Almquist, DuBois
  - ACM SIGKDD Conference (DuBois)
  - International Conference on Machine Learning (Vu)
  - American Sociological Association Meeting (Marcum, Jasny, Spiro, Fitzhugh, Almquist)
- Best paper awards or nominations (Spiro, Hummel)
- National fellowships: DuBois (NDSEG), Asuncion (NSF), Navaroli (NDSEG)

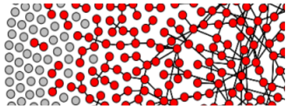
…..and the Old Generation

- Carter Butts
  - American Sociological Association, Leo A. Goodman award, 2010
  - highest award to young methodological researchers in social science

- Michael Goodrich
  - ACM Fellow, IEEE Fellow, 2009

- Padhraic Smyth
  - ACM SIGKDD Innovation Award 2009
  - AAAI Fellow 2010

- Mark Handcock
  - Fellow of the American Statistical Association, 2009

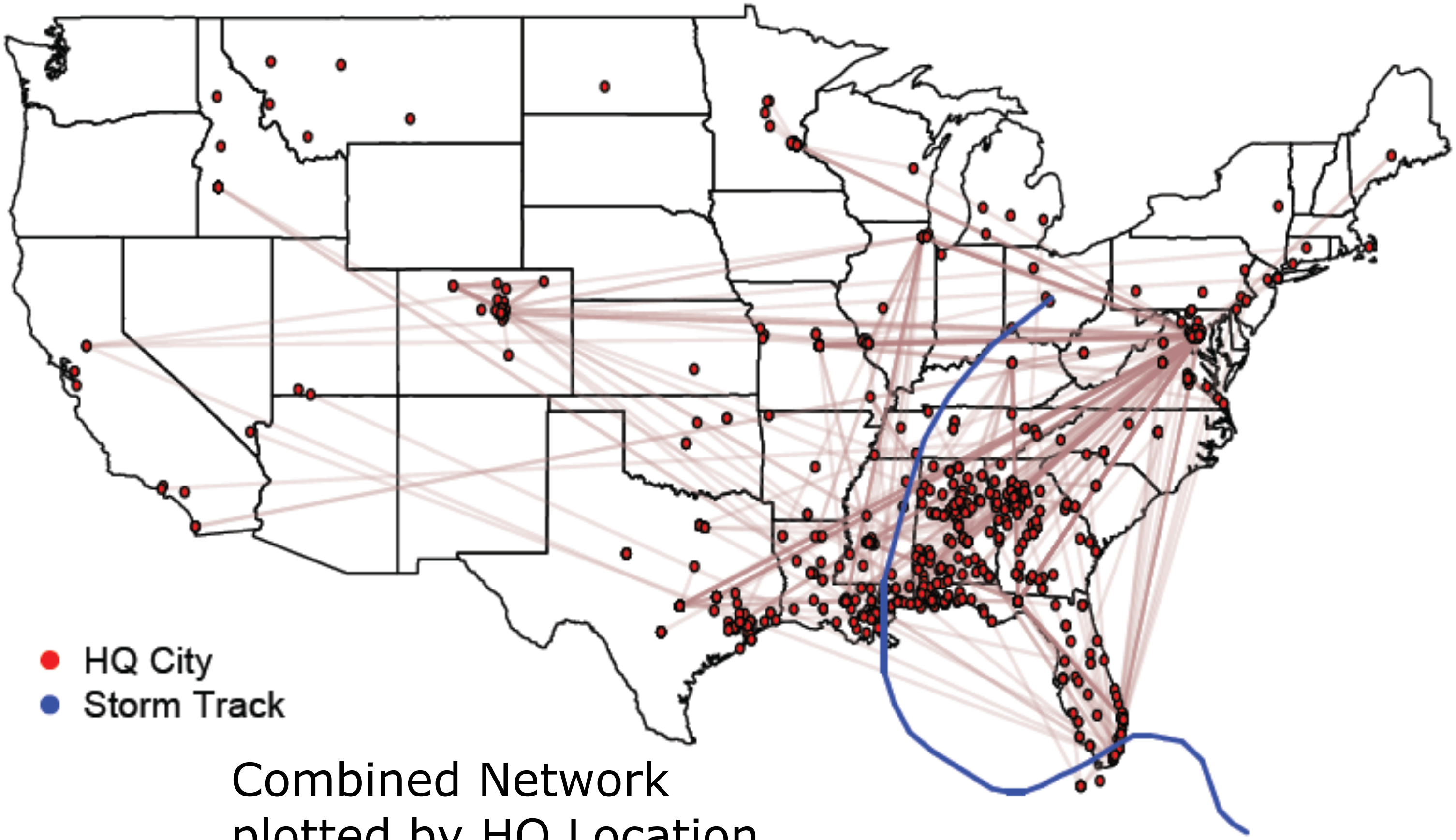


What Next?

- “Push” algorithmic advances into statistical modeling
  - Will allow us to scale existing algorithms to much larger data sets
- Develop network models with richer representational power
  - Geographic data, temporal events, text data, actor covariates, heterogeneity, etc
- Systematically evaluate and test different approaches
  - evaluate ability of models to predict over time, to impute missing values, etc
- Apply these approaches to high visibility problems and data sets
  - E.g., online social interaction such as email, Facebook, Twitter, blogs
- Make software publicly available

Organizational Collaboration during the Katrina Disaster

Almquist and Butts



Combined Network plotted by HQ Location

SESSION 1:

9:20 Dynamic Egocentric Models for Citation Networks
Dave Hunter, Professor, Statistics, Penn State University

9:45 Membership Dimension
Maarten Loffler, Postdoctoral Fellow, Computer Science, UC Irvine

10:05 Multilevel Network Models for Classroom Dynamics
Chris DuBois, PhD student, Statistics, UC Irvine

10:30 COFFEE BREAK

SESSION 2:

10:45 DISCUSSION: FAST CHANGE-SCORE COMPUTATION IN DYNAMIC GRAPHS
Led by David Eppstein and Michael Goodrich, Computer Science, UC Irvine

11:35 Bayesian Meta-Analysis of Network Data via Reference Quantiles
Carter Butts, Professor, Social Sciences, UC Irvine

12:00 Break for lunch (lunch for PIs + visitors at the University Club)

1:30 to 2:40 Short Highlight Talks

Computational Issues with Exponential Random Graph Models
Mark Handcock, Professor, Statistics, UCLA

Experimental Results on Fast Clique Finding
David Eppstein, Professor, Computer Science, UC Irvine

Modeling Rates between Affiliates on Facebook from Sampled Data
Emma Spiro, PhD student, Social Sciences, UC Irvine

Modeling Degree Sequences of Undirected Networks with Application to 9/11 Disaster Networks
Miruna Petrescu-Prahova, Postdoctoral Fellow, Statistics, University of Washington

Statistical Models for Text and Networks
Jimmy Foulds, PhD student, Computer Science, UC Irvine

Analysis of Life History Data
Sean Fitzhugh, PhD student, Social Sciences, UC Irvine

Approximate Sampling for Binary Discrete Exponential Families with Fixed Execution Time and Quality Guarantees
Carter Butts, Professor, Social Sciences, UC Irvine

2:40 – 3:30: SESSION 3

2:40 Instability, Sensitivity, and Degeneracy of Discrete Exponential Families
Michael Schweinberger, Postdoctoral Fellow, Penn State University

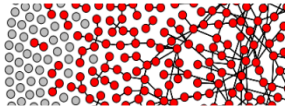
3:05 Empirical Analysis of Latent Space Embedding
David Mount, Professor, University of Maryland

3:30 COFFEE BREAK

4:00 DISCUSSION: LATENT VARIABLE MODELING OF NETWORK DATA
Led by Carter Butts and Padhraic Smyth

4:45 WRAP-UP, CLOSING COMMENTS

5:00 ADJOURN



Additional Resources

Project Web site:

http://www.datalab.uci.edu/muri/

Slides and Posters from AHM:

http://www.datalab.uci.edu/muri/june2011/

Publications:

http://www.datalab.uci.edu/muri/publications.php

Software:

http://csde.washington.edu/statnet/

Data Sets:

http://networkdata.ics.uci.edu/resources.php

QUESTIONS?