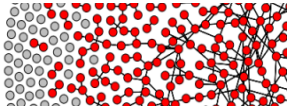
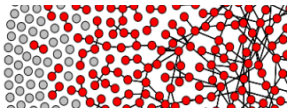


Scalable Methods for the Analysis of Network-Based Data

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

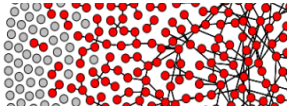
Slides online at www.datalab.uci.edu/muri



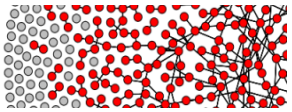
Today's Meeting

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

- Format
  - Introduction
  - Research talks
    - 25 minute slots
    - 5 mins at end for questions/discussion
  - Poster session from 1:15 to 2:45
  - Question/discussion encouraged during talks
  - Several breaks for discussion

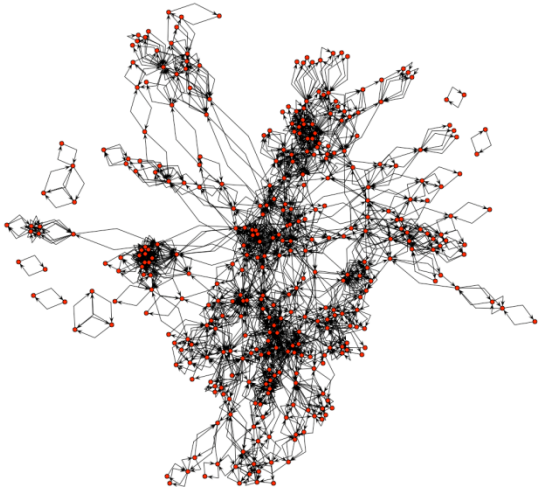


Motivation and Background



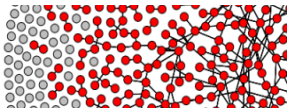
Motivation

2007: interdisciplinary interest in analysis of large network data sets



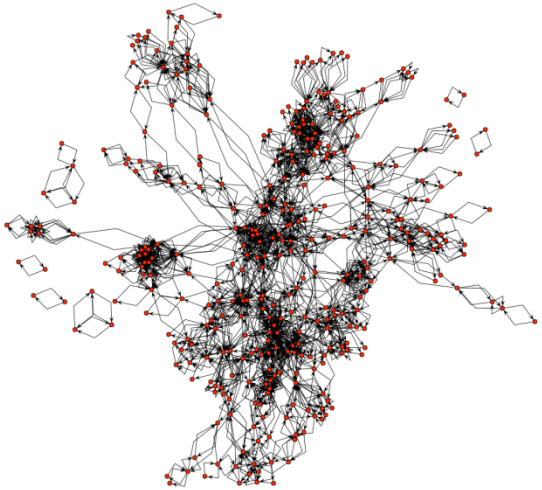
Many of the available techniques are descriptive, cannot handle

- Prediction
- Missing data
- Covariates, etc



Motivation

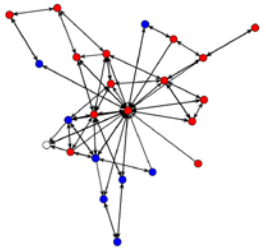
2007: interdisciplinary interest in analysis of large network data sets



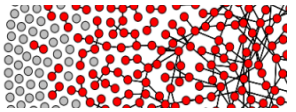
Many of the available techniques are descriptive, cannot handle

- Prediction
- Missing data
- Covariates, etc

2007: significant statistical body of theory available on network modeling

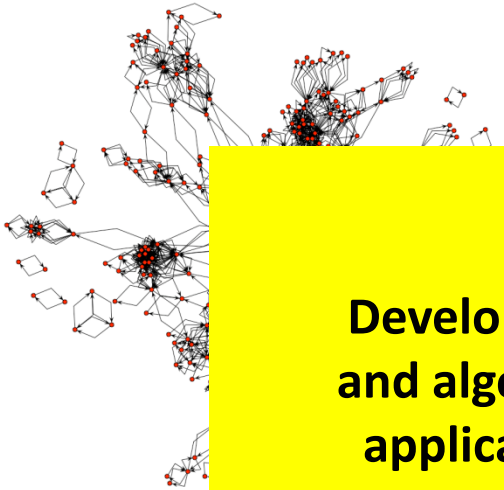


Many of the available techniques do not scale up to large data sets, not widely known/understood/used, etc

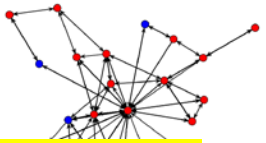


Motivation

2007: interdisciplinary interest in analysis of large network data sets



2007: significant statistical body of theory available on network modeling



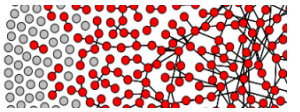
Goal of this MURI project

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

chniques do not
a sets, not widely
used, etc

Many of the avai
descriptive,

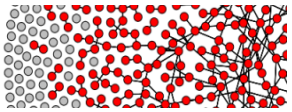
- Prediction
- Missing data
- Covariates, etc



Project Dates

- Project Timeline
  - Start date: May 1 2008
  - End date: April 30 2011 (for 3-year award)

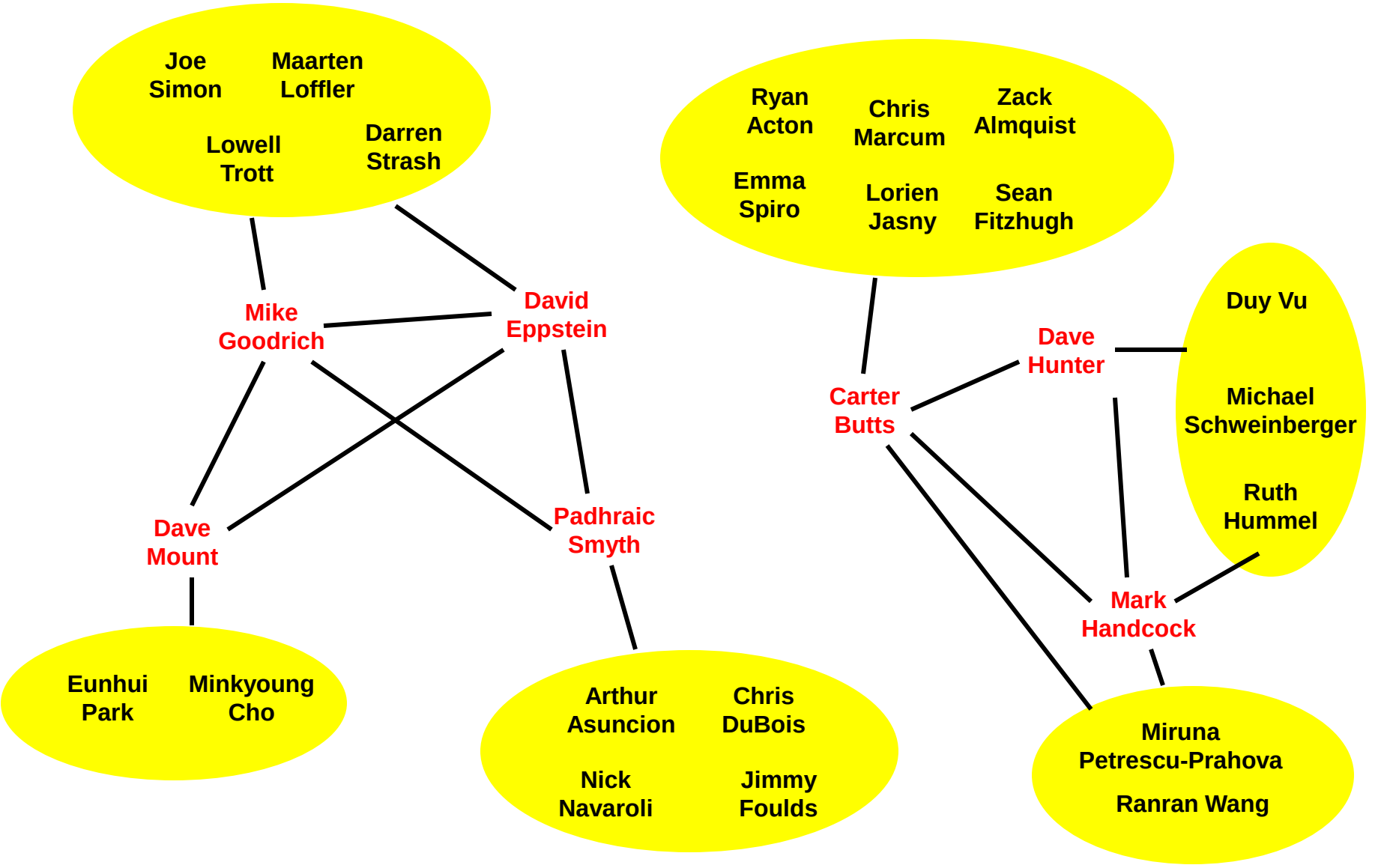
- Meetings
  - Kickoff Meeting, November 2008
  - Working Meeting, April 2009
  - Working Meeting, August 2009
  - Annual Review, December 2009
  - Working Meeting, May 2010
  - Annual Review, November 2010



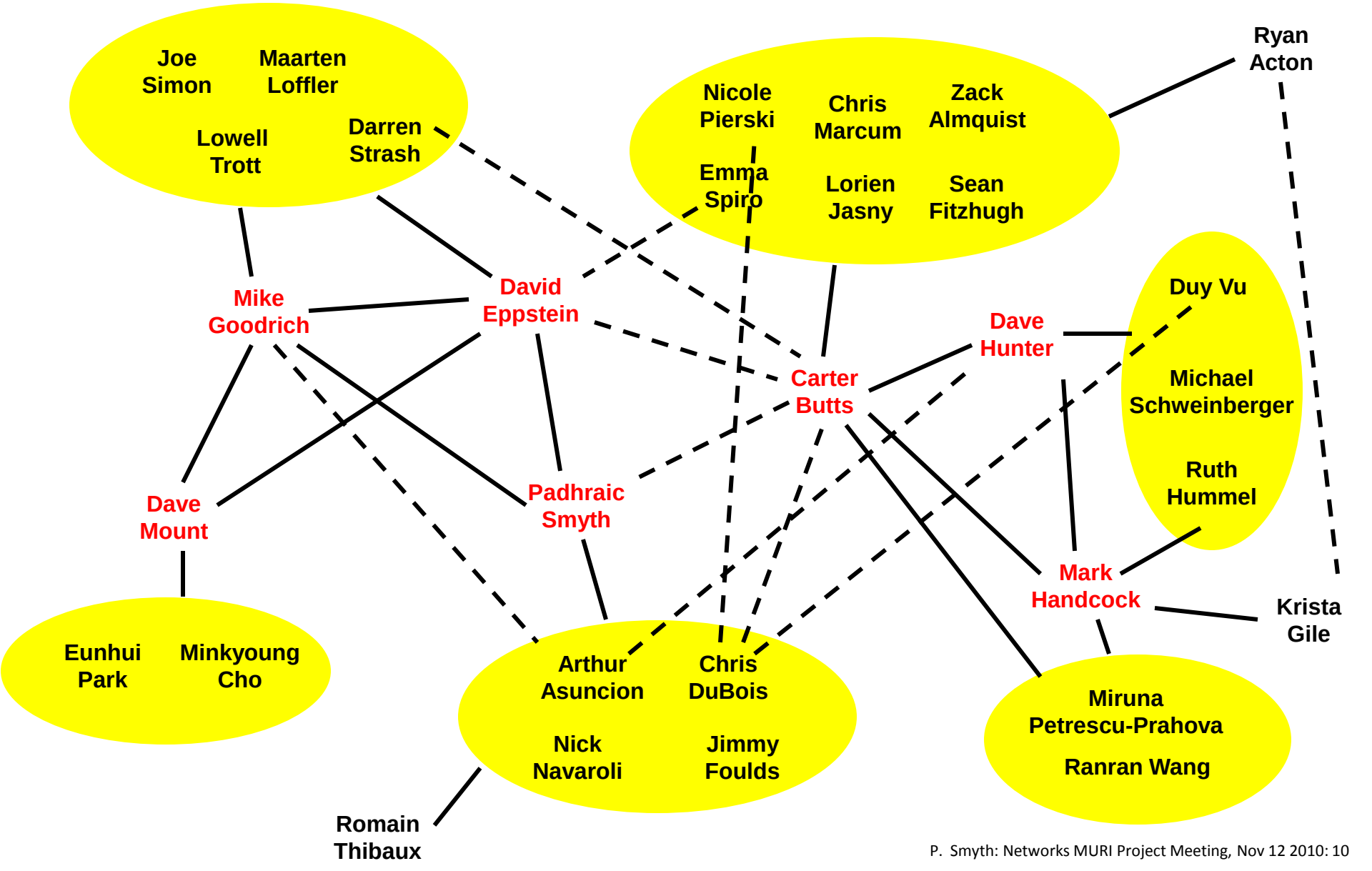
MURI Team

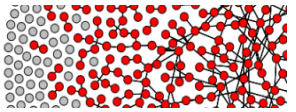
Investigator	University	Department(s)	Expertise	Number Of PhD Students	Number of Postdocs
 Padhraic Smyth (PI)	UC Irvine	Computer Science	Machine learning	4	
 Carter Butts	UC Irvine	Sociology	Statistical social network analysis	6	
 Mark Handcock	UCLA	Statistics	Statistical social network analysis	1	1
 Dave Hunter	Penn State	Statistics	Computational statistics	2	1
 David Eppstein	UC Irvine	Computer Science	Graph algorithms	2	1
 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**	**4**

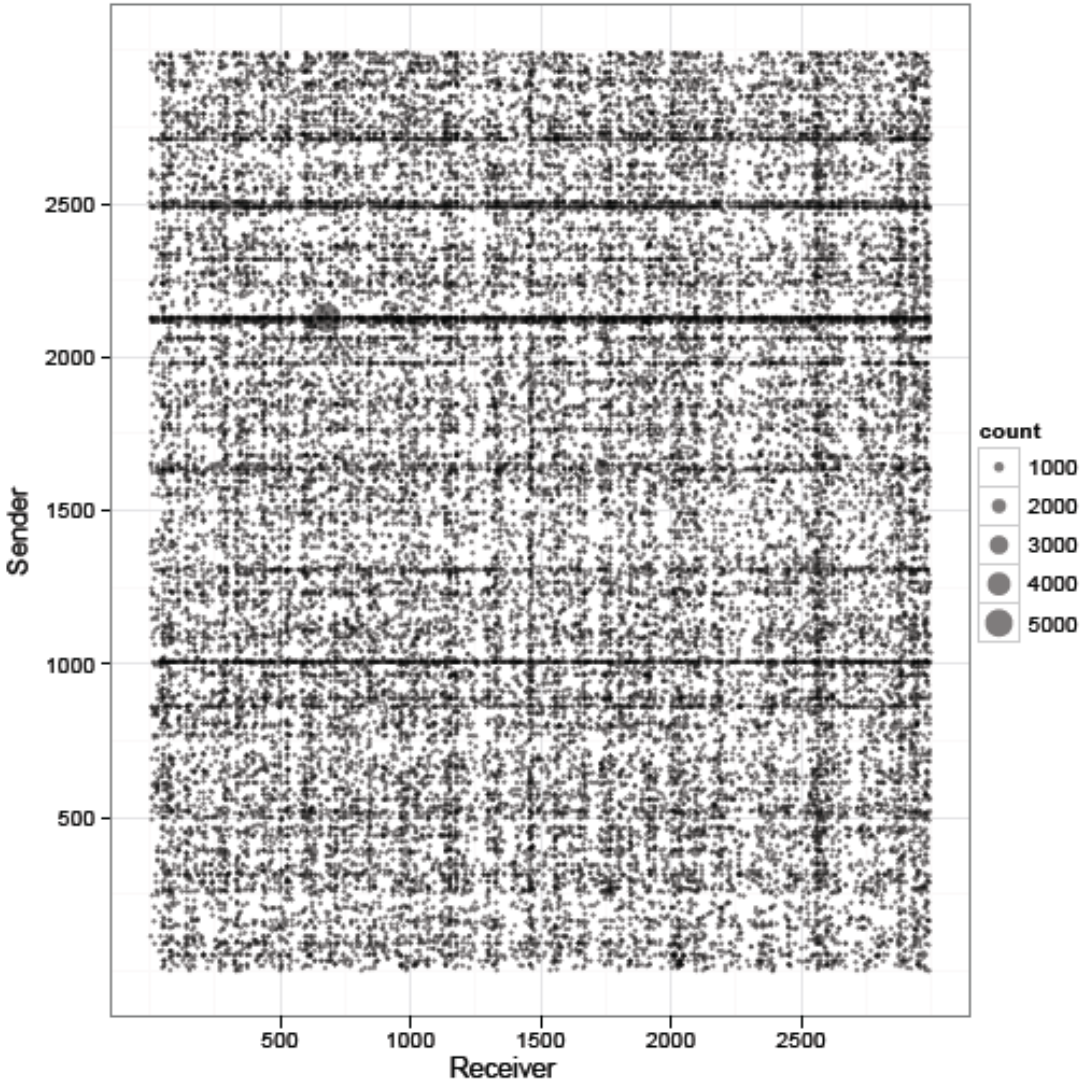
Collaboration Network



Collaboration Network



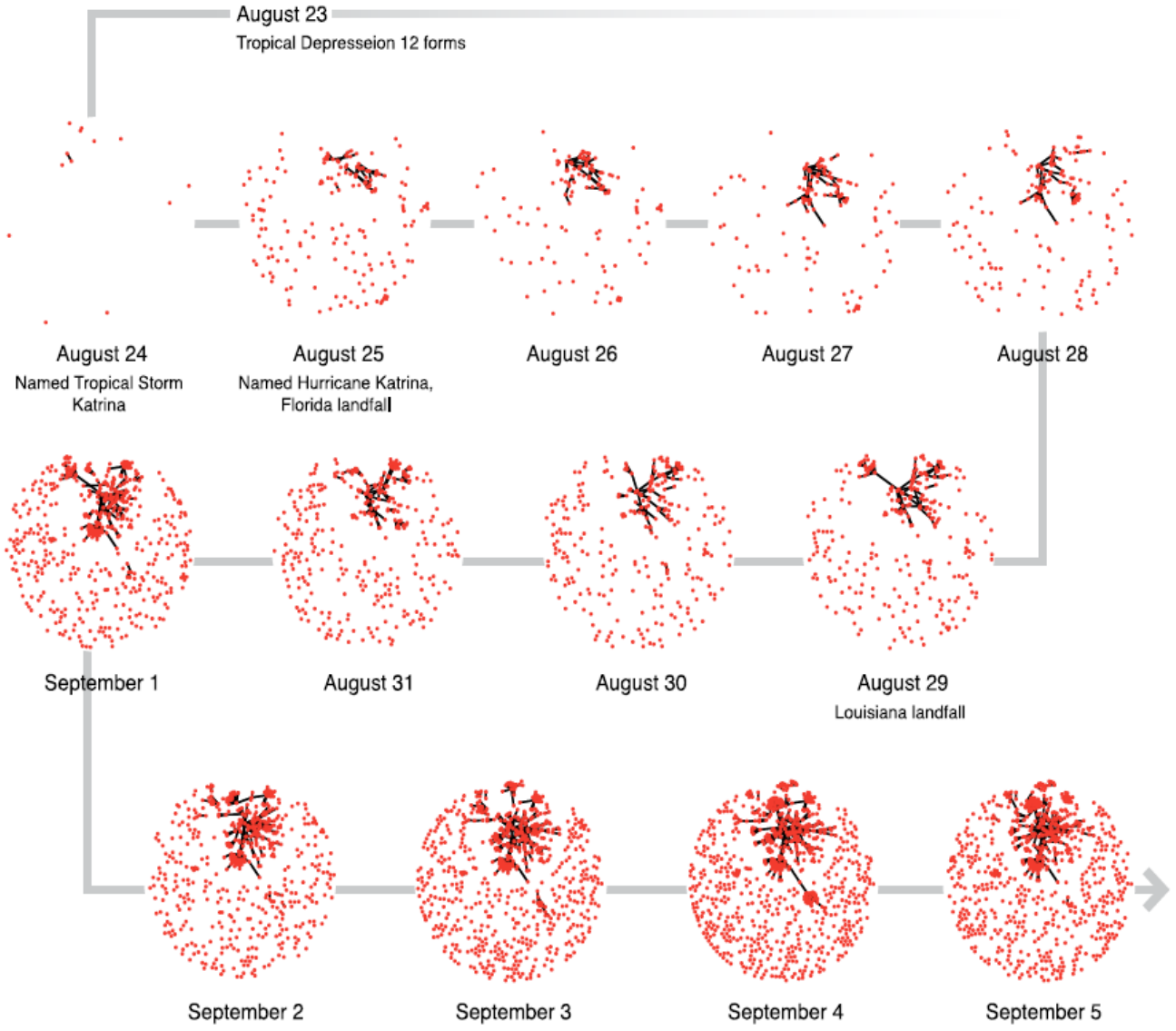




Data:
Count matrix of 200,000 email messages among 3000 individuals over 3 months

Problem:
Understand communication pattterns and predict future communication activity

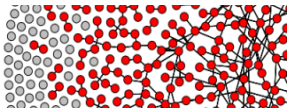
Challenges:
sparse data, missing data, non-stationarity, unseen covariates



Data:
Inter-organizational communication patterns over time, post-Katrina

Problem:
understand the processes underlying network growth

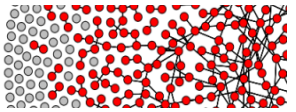
Challenge:
noisy and sparse data, missing covariates



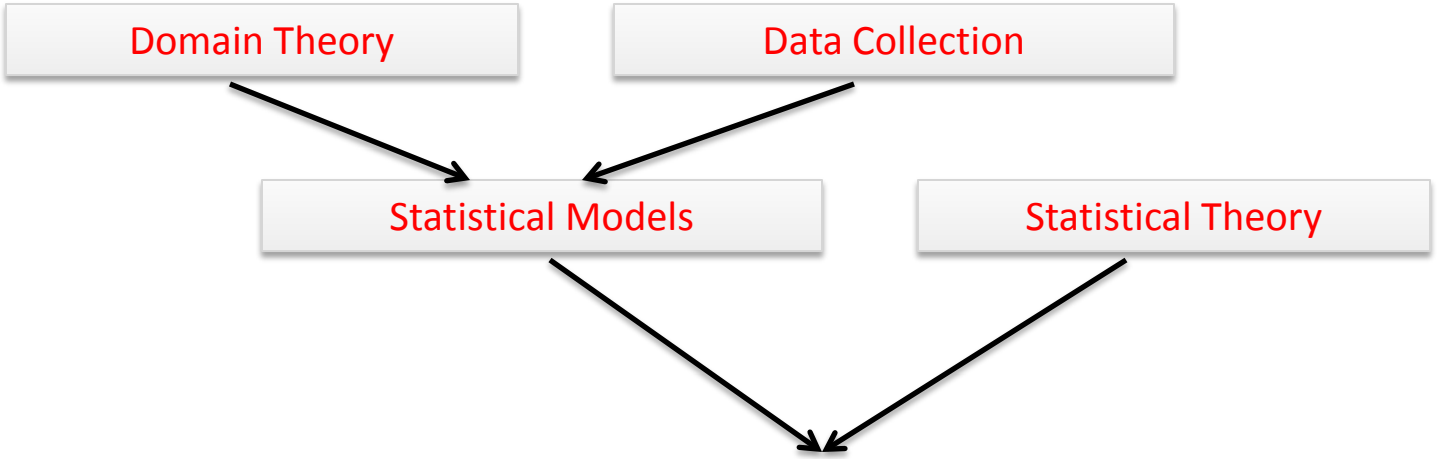
Key Scientific/Technical Challenges

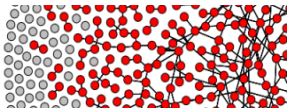
- Parametrize models in a sensible and computable way
  - Respect theories of social behavior as well as explain observed data, in a computationaly scalable manner
- Account for real data
  - Understand sampling methods: account for missing, error-prone data
- Make inference both principled and practical
  - 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

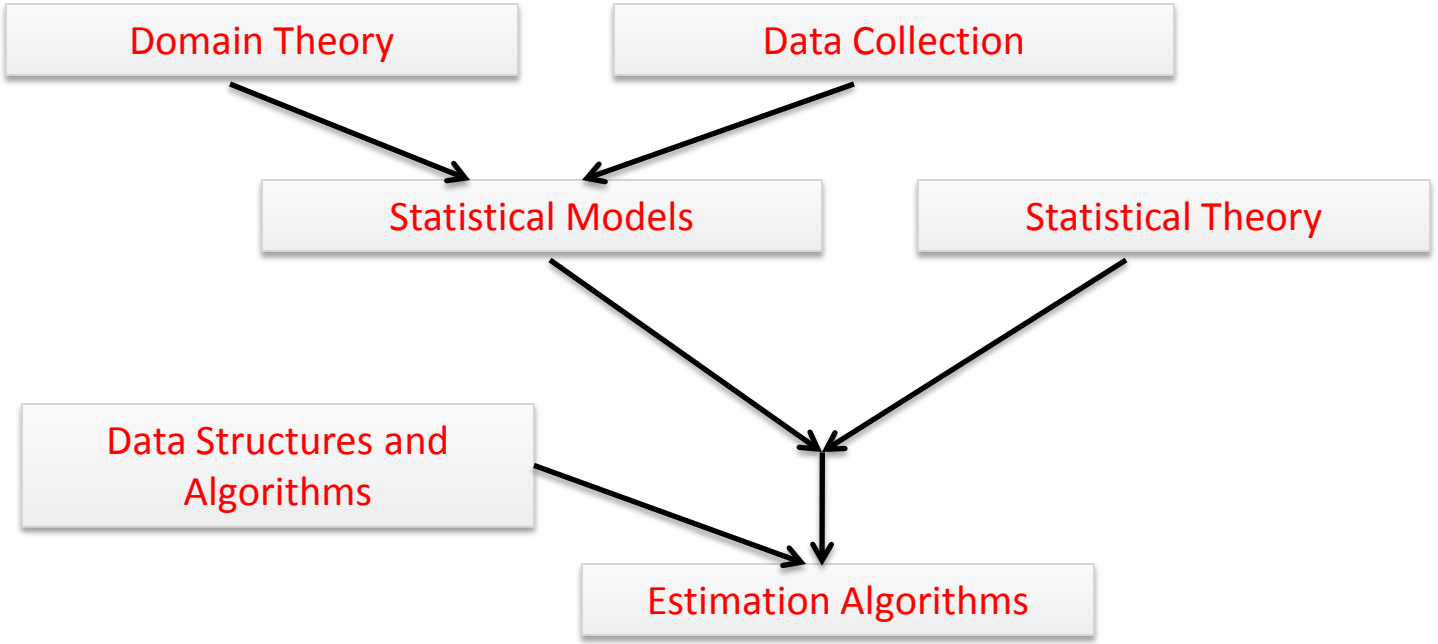


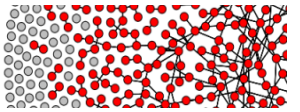
Mapping the Project Terrain



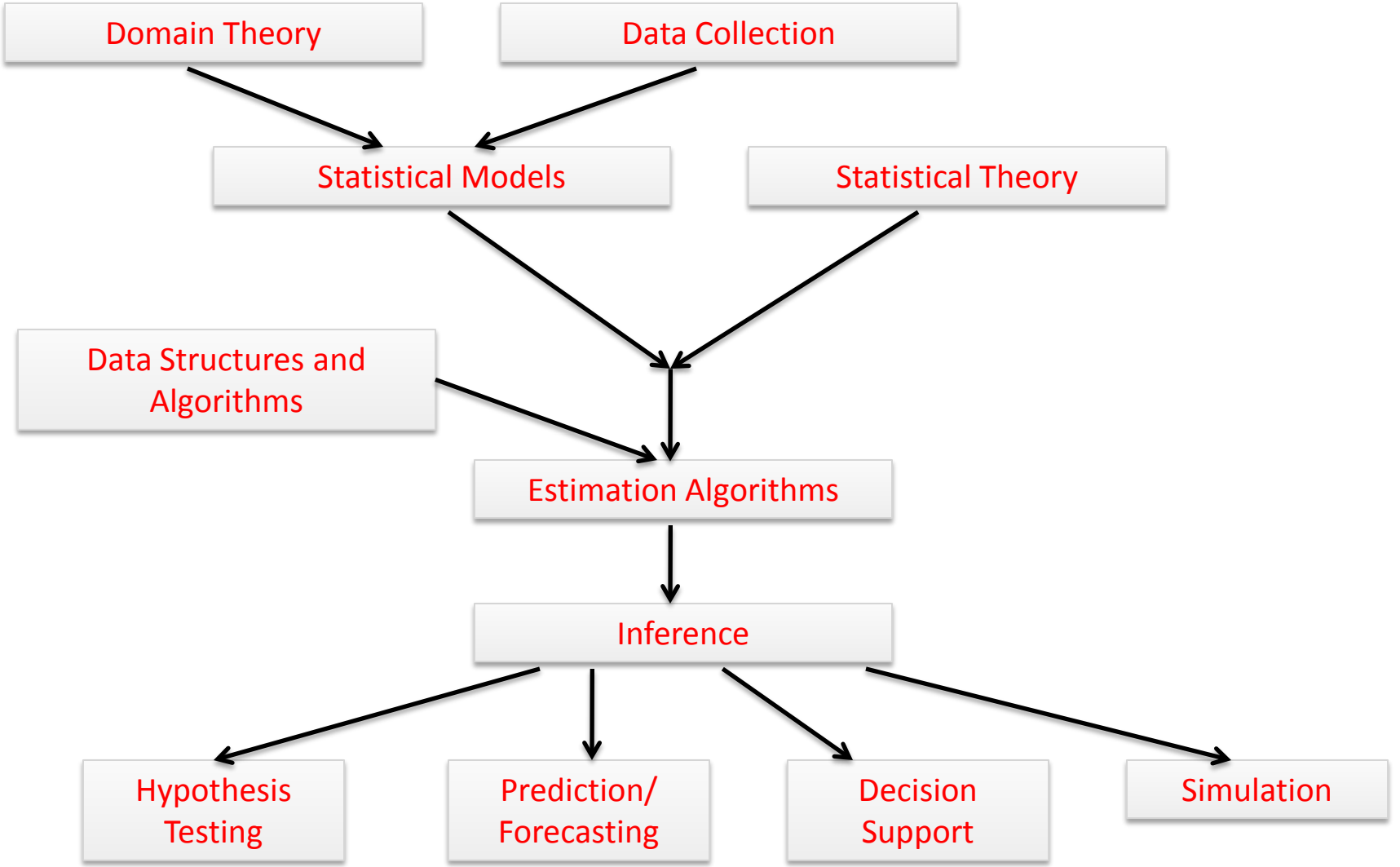


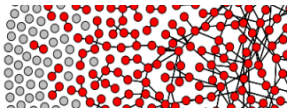
Mapping the Project Terrain





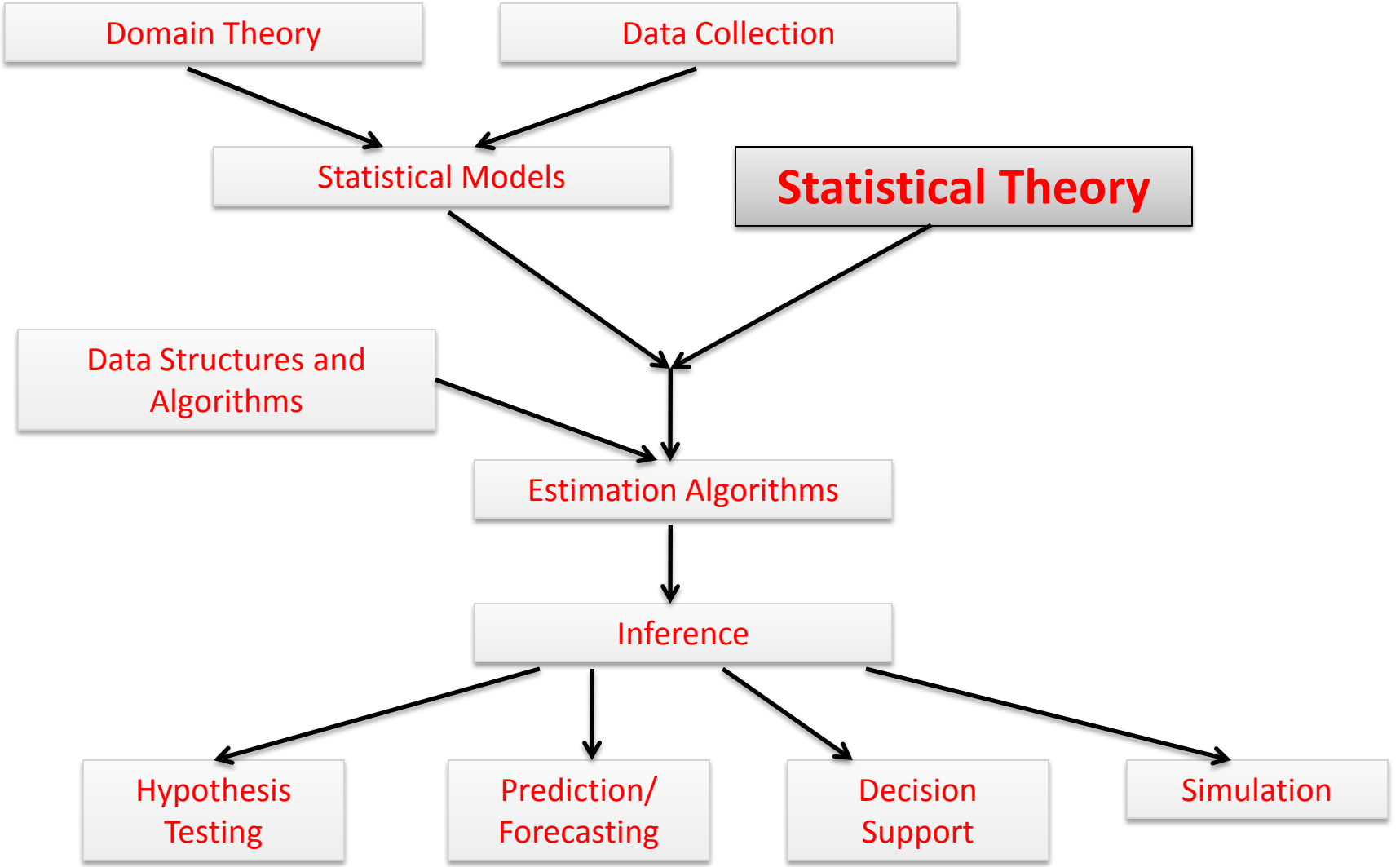
Mapping the Project Terrain

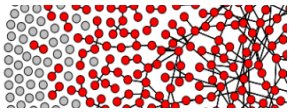




Summary of Accomplishments

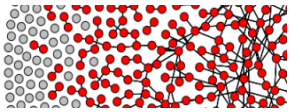
Mapping the Project Terrain





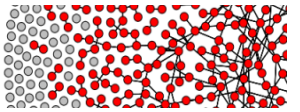
Accomplishments: Theory and Methodology

Topic	State of the Art in 2008	State of the Art now (with MURI)	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
Hidden/network population sampling	No method for assessing sample quality No method for sampling with no well-connected network	New principled methods for assessing convergence. New multigraph sampling for non-connected networks (Butts el al, 2010)	Potentially significant new applications in areas such as criminology, epidemiology, etc
Theory for complex network models	Little theory for non-Bernoulli models – knowledge based on approximate simulations	New method based on "Bernoulli graph bounds" (Butts, 2009)	Tools for understanding of model properties will allow us to focus on better models



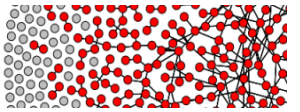
Accomplishments: Theory and Methodology

Topic	State of the Art in 2008	State of the Art now (with MURI)	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
Hidden/network population sampling	No method for assessing sample quality No method for sampling with no well-connected network	New principled methods for assessing convergence. New multigraph sampling for non-connected networks (Butts el al, 2010)	Potentially significant new applications in areas such as criminology, epidemiology, etc
Theory for complex network models	Little theory for non-Bernoulli models – knowledge based on approximate simulations	New method based on "Bernoulli graph bounds" (Butts, 2009)	Tools for understanding of model properties will allow us to focus on better models



Accomplishments: Theory and Methodology

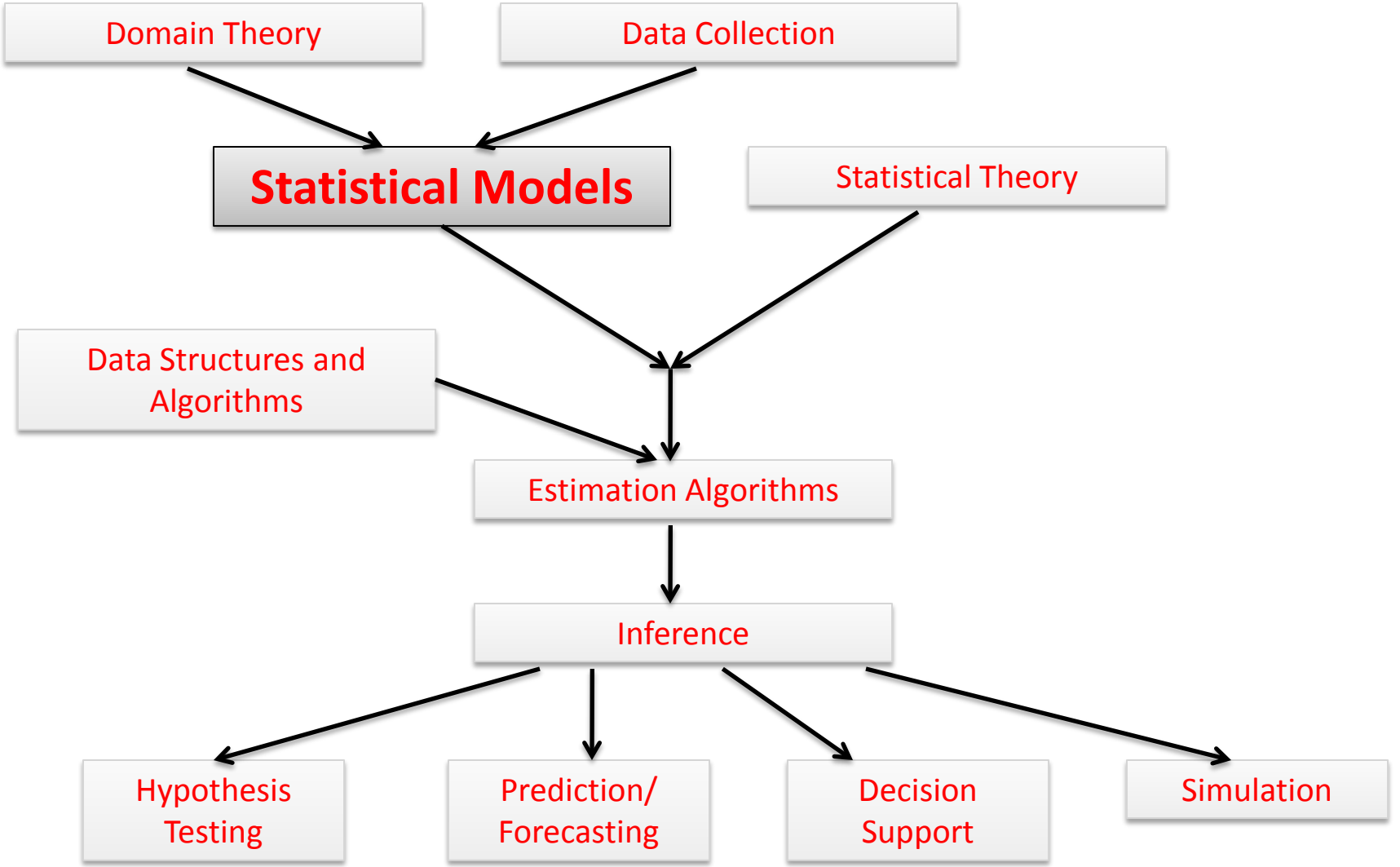
Topic	State of the Art in 2008	State of the Art now (with MURI)	Potential Applications And Impact
General theory for handling missing data in social networks	Problem only partially understood. No software available for statistical modeling	General statistical theory for treating missing data in a social network context. Publicly-available code in R. (Gile and Handcock, 2010)	Allows application of social network modeling to data sets with significant missing data
Hidden/network population sampling	No method for assessing sample quality No method for sampling with no well-connected network	New principled methods for assessing convergence. New multigraph sampling for non-connected networks (Butts el al, 2010)	Potentially significant new applications in areas such as criminology, epidemiology, etc
Theory for complex network models	Little theory for non-Bernoulli models – knowledge based on approximate simulations	New method based on "Bernoulli graph bounds" (Butts, 2009)	Tools for understanding of model properties will allow us to focus on better models

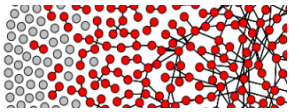


Accomplishments: Theory and Methodology

Topic	State of the Art in 2008	State of the Art now (with MURI)	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
Hidden/network population sampling	No method for assessing sample quality No method for sampling with no well-connected network	New principled methods for assessing convergence. New multigraph sampling for non-connected networks (Butts el al, 2010)	Potentially significant new applications in areas such as criminology, epidemiology, etc
Theory for complex network models	Little theory for non-Bernoulli models – knowledge based on approximate simulations	New method based on "Bernoulli graph bounds" (Butts, 2009)	Tools for understanding of model properties will allow us to focus on better models

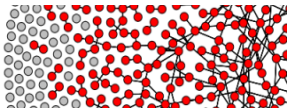
Mapping the Project Terrain





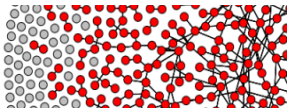
Accomplishments: Network Data over Time

Topic	State of the Art in 2008	State of the Art now (with MURI)	Potential Applications And Impact
Modeling of network dynamics	100 nodes, 10 time points (e.g., SIENA package)	1000's of nodes, 1000's of time points Based on logistic approximation (Almquist and Butts, 2010)	Expands applicability of dynamic network modeling to large realistic applications, as well as scope of questions that can be addressed
Relational event models	Basic dyadic event models. No exogenous events. No public software.	Much richer model with exogenous events, egocentric support, multiple observer accounts (Butts et al, 2010)	
Imputing missing events in dynamic network data	No general purpose method published No software available	Accurate and computationally efficient imputation using latent class models Software publicly available (DuBois and Smyth, 2010)	



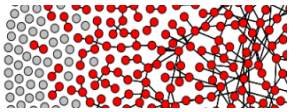
Accomplishments: Network Data over Time

Topic	State of the Art in 2008	State of the Art now (with MURI)	Potential Applications And Impact
Modeling of network dynamics	100 nodes, 10 time points (e.g., SIENA package)	1000's of nodes, 1000's of time points Based on logistic approximation (Almquist and Butts, 2010)	Expands applicability of dynamic network modeling to large realistic applications, as well as scope of questions that can be addressed
Relational event models	Basic dyadic event models. No exogenous events. No public software.	Much richer model with exogenous events, egocentric support, multiple observer accounts (Butts et al, 2010)	
Imputing missing events in dynamic network data	No general purpose method published No software available	Accurate and computationally efficient imputation using latent class models Software publicly available (DuBois and Smyth, 2010)	



Accomplishments: Network Data over Time

Topic	State of the Art in 2008	State of the Art now (with MURI)	Potential Applications And Impact
Modeling of network dynamics	100 nodes, 10 time points (e.g., SIENA package)	1000's of nodes, 1000's of time points Based on logistic approximation (Almquist and Butts, 2010)	Expands applicability of dynamic network modeling to large realistic applications, as well as scope of questions that can be addressed
Relational event models	Basic dyadic event models. No exogenous events. No public software.	Much richer model with exogenous events, egocentric support, multiple observer accounts (Butts et al, 2010)	
Imputing missing events in dynamic network data	No general purpose method published No software available	Accurate and computationally efficient imputation using latent class models Software publicly available (DuBois and Smyth, 2010)	

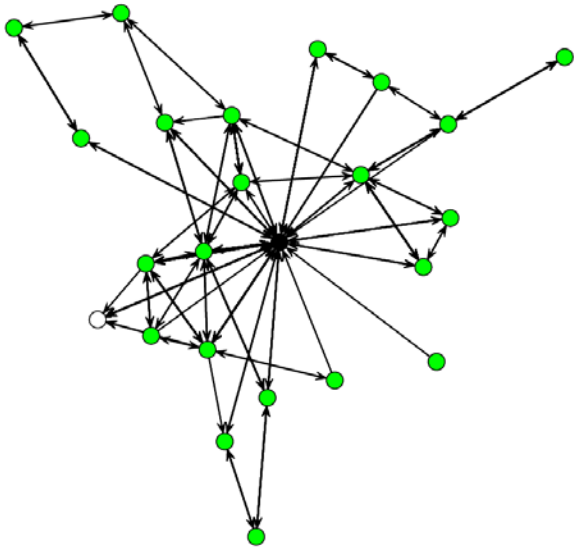


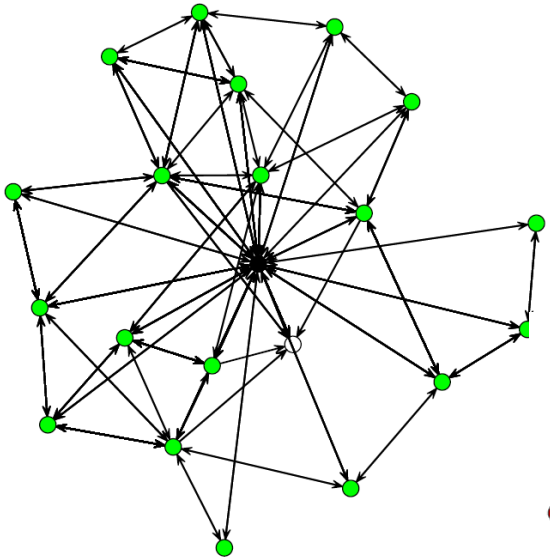
Accomplishments: Network Data over Time

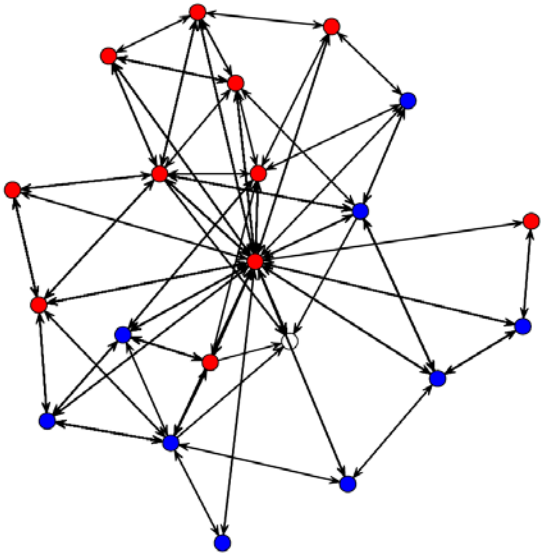
Topic	State of the Art in 2008	State of the Art now (with MURI)	Potential Applications And Impact
Modeling of network dynamics	100 nodes, 10 time points (e.g., SIENA package)	1000's of nodes, 1000's of time points Based on logistic approximation (Almquist and Butts, 2010)	Expands applicability of dynamic network modeling to large realistic applications, as well as scope of questions that can be addressed
Relational event models	Basic dyadic event models. No exogenous events. No public software.	Much richer model with exogenous events, egocentric support, multiple observer accounts (Butts et al, 2010)	
Imputing missing information in dynamic network data	No general purpose method published No software available	Accurate and computationally efficient imputation using latent class models Software publicly available (DuBois and Smyth, 2010)	

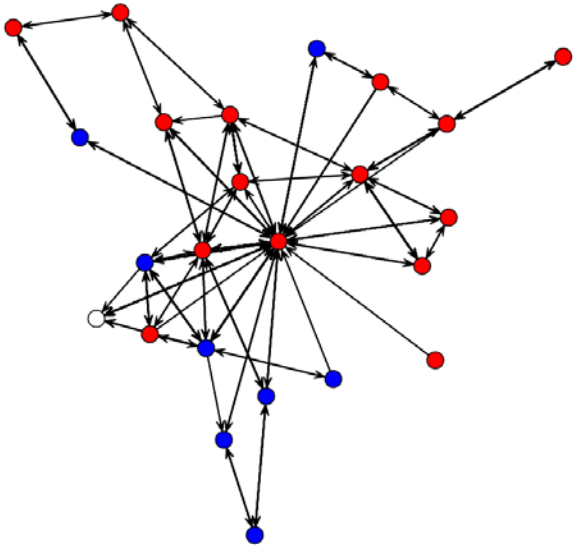
Network Dynamics in Classroom Interactions

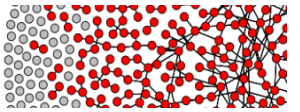
Poster by PhD student Nicole Pierski



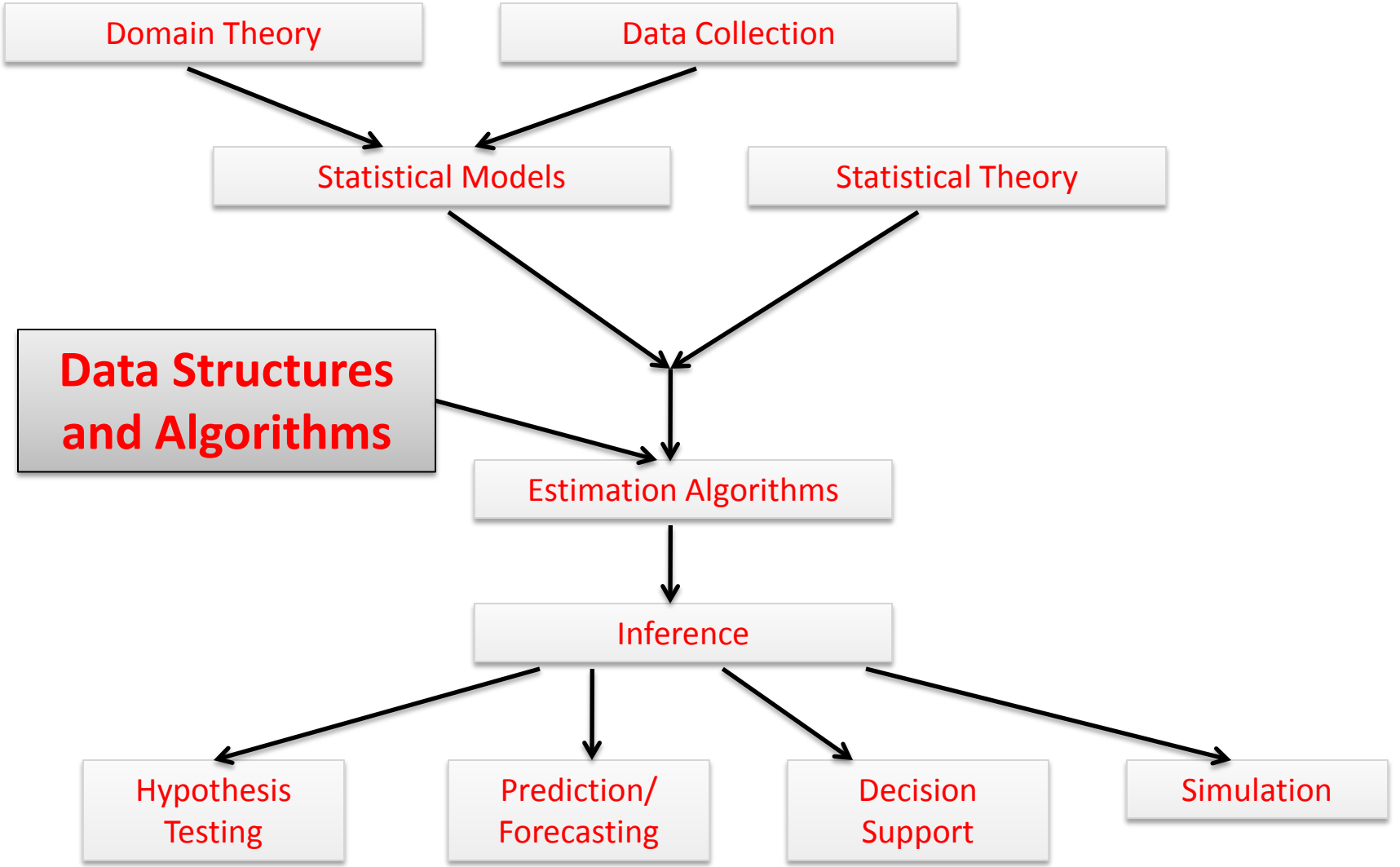


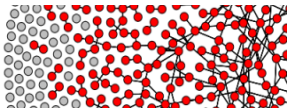






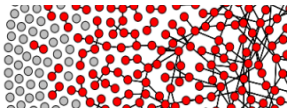
Mapping the Project Terrain





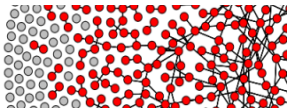
Accomplishments: Data Structures and Algorithms

Topic	State of the Art in 2008	State of the Art now (with MURI)	Potential Applications And Impact
Dynamically-changing graphs	Dynamic graph algorithms not applied to social network modeling	Efficient new algorithms for dynamically maintaining counts of ERGM features (Eppstein and Spiro, 2009; Eppstein et al, 2010)	Extends applicability of statistical network modeling to larger networks and more complex models
Latent space computations	Learning algorithm scales poorly: each iteration is quadratic in N	New more efficient algorithms based on geometric data structures (Mount and Park 2010)	
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)	



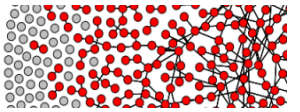
Accomplishments: Data Structures and Algorithms

Topic	State of the Art in 2008	State of the Art now (with MURI)	Potential Applications And Impact
Dynamically-changing graphs	Dynamic graph algorithms not applied to social network modeling	Efficient new algorithms for dynamically maintaining counts of ERGM features (Eppstein and Spiro, 2009; Eppstein et al, 2010)	Extends applicability of statistical network modeling to larger networks and more complex models
Latent space computations	Learning algorithm scales poorly: each iteration is quadratic in N	New more efficient algorithms based on geometric data structures (Mount and Park 2010)	
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)	



Accomplishments: Data Structures and Algorithms

Topic	State of the Art in 2008	State of the Art now (with MURI)	Potential Applications And Impact
Dynamically-changing graphs	Dynamic graph algorithms not applied to social network modeling	Efficient new algorithms for dynamically maintaining counts of ERGM features (Eppstein and Spiro, 2009; Eppstein et al, 2010)	Extends applicability of statistical network modeling to larger networks and more complex models
Latent space computations	Learning algorithm scales poorly: each iteration is quadratic in N	New more efficient algorithms based on geometric data structures (Mount and Park 2010)	
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)	

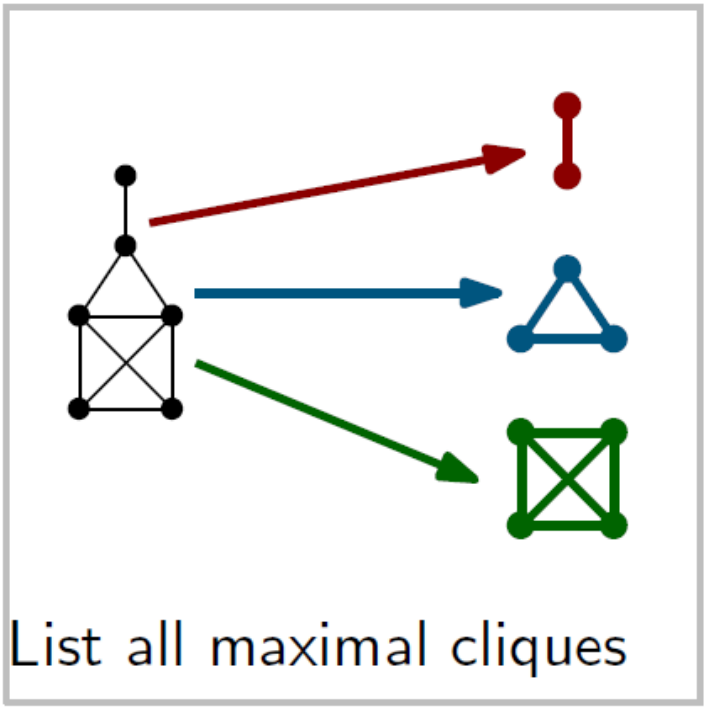


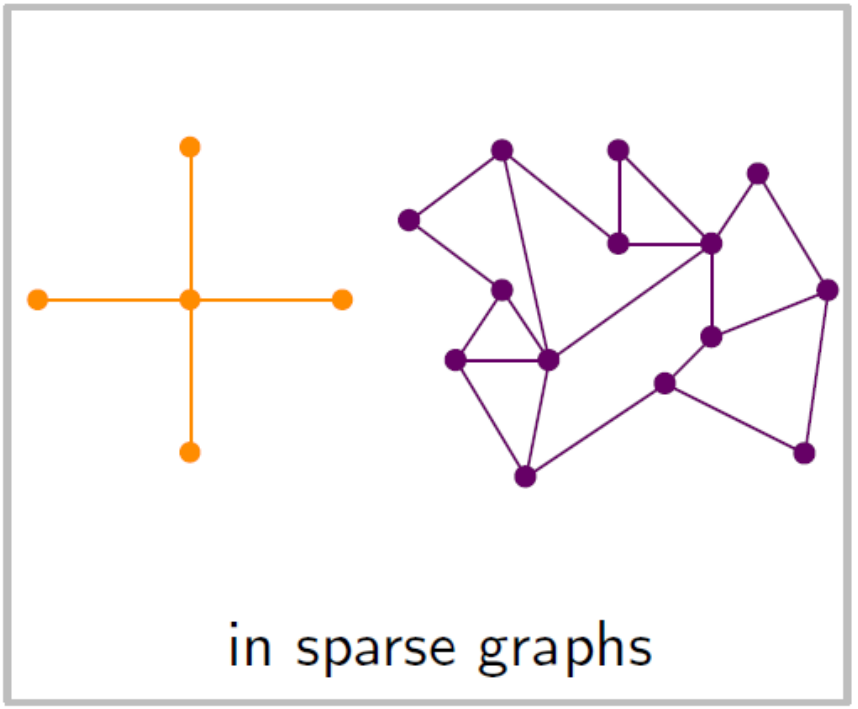
Accomplishments: Data Structures and Algorithms

Topic	State of the Art in 2008	State of the Art now (with MURI)	Potential Applications And Impact
Dynamically-changing graphs	Dynamic graph algorithms not applied to social network modeling	Efficient new algorithms for dynamically maintaining counts of ERGM features (Eppstein and Spiro, 2009; Eppstein et al, 2010)	Extends applicability of statistical network modeling to larger networks and more complex models
Latent space computations	Learning algorithm scales poorly: each iteration is quadratic in N	New more efficient algorithms based on geometric data structures (Mount and Park 2010)	
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)	

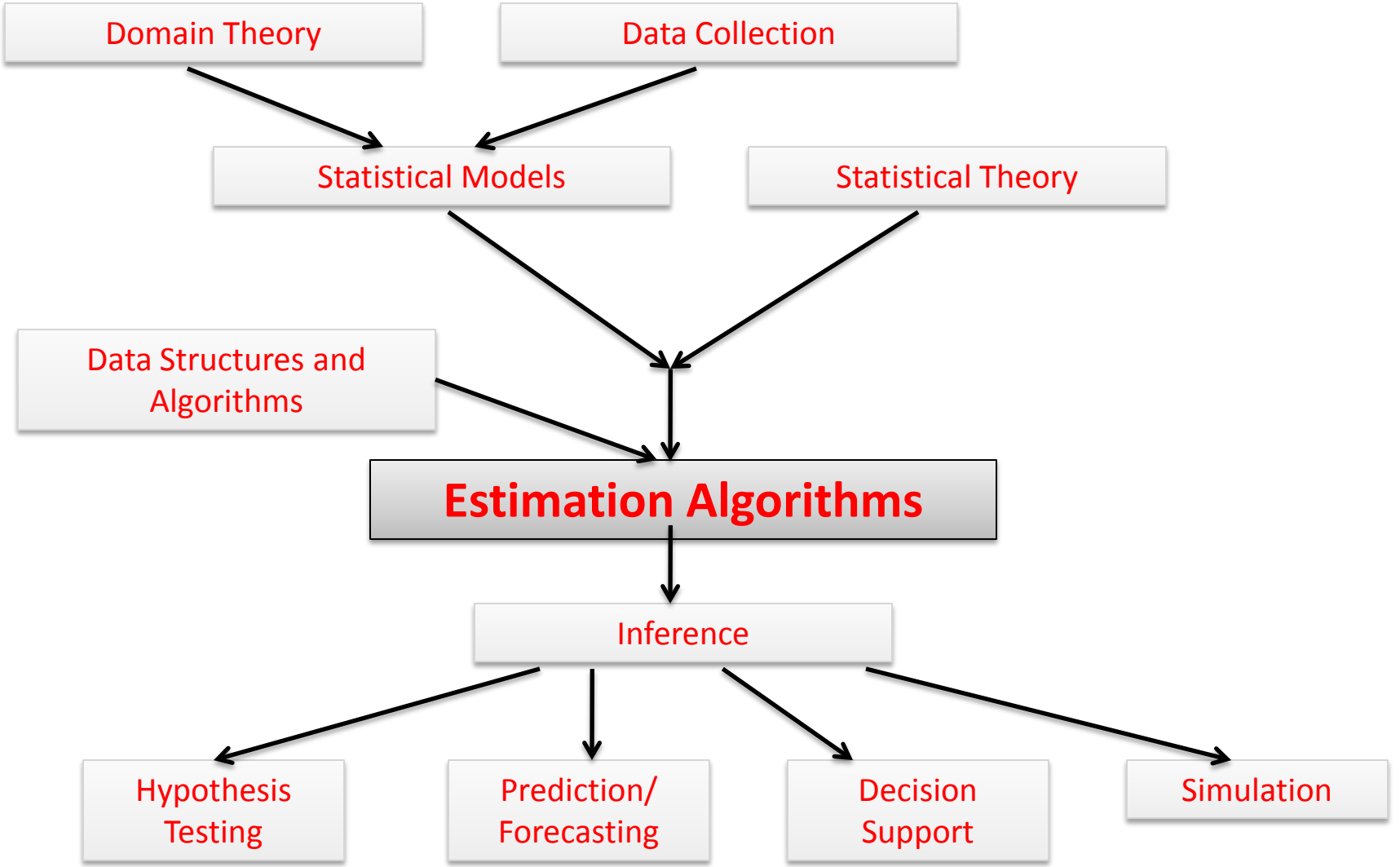
Finding All Maximal Cliques in Sparse Graphs

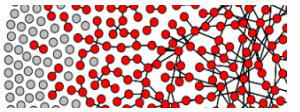
Talk by PhD student Darren Strash





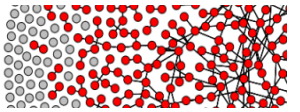
Mapping the Project Terrain





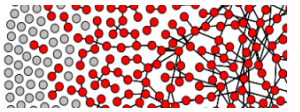
Accomplishments: Scalability

Topic	State of the Art in 2008	State of the Art now (with MURI)	Potential Applications And Impact
Mixtures of ERG models	600 nodes binary-valued (Daudin et al, 2008)	100,000 nodes Categorical-valued (Hunter and Vu, 2010)	Broadens applicability of statistical inference to large noisy networks
Latent variable network models	100 nodes (Raftery et al, JRSS, 2006)	100,000 nodes Efficient latent-class algorithm (DuBois and Smyth, 2010)	Extends statistical network models to data sets where only descriptive methods were used previously



Accomplishments: Scalability

Topic	State of the Art in 2008	State of the Art now (with MURI)	Potential Applications And Impact
Mixtures of ERG models	600 nodes binary-valued (Daudin et al, 2008)	100,000 nodes Categorical-valued (Hunter and Vu, 2010)	Broadens applicability of statistical inference to large noisy networks
Latent variable network models	100 nodes (Raftery et al, JRSS, 2006)	100,000 nodes Efficient latent-class algorithm (DuBois and Smyth, 2010)	Extends statistical network models to data sets where only descriptive methods were used previously

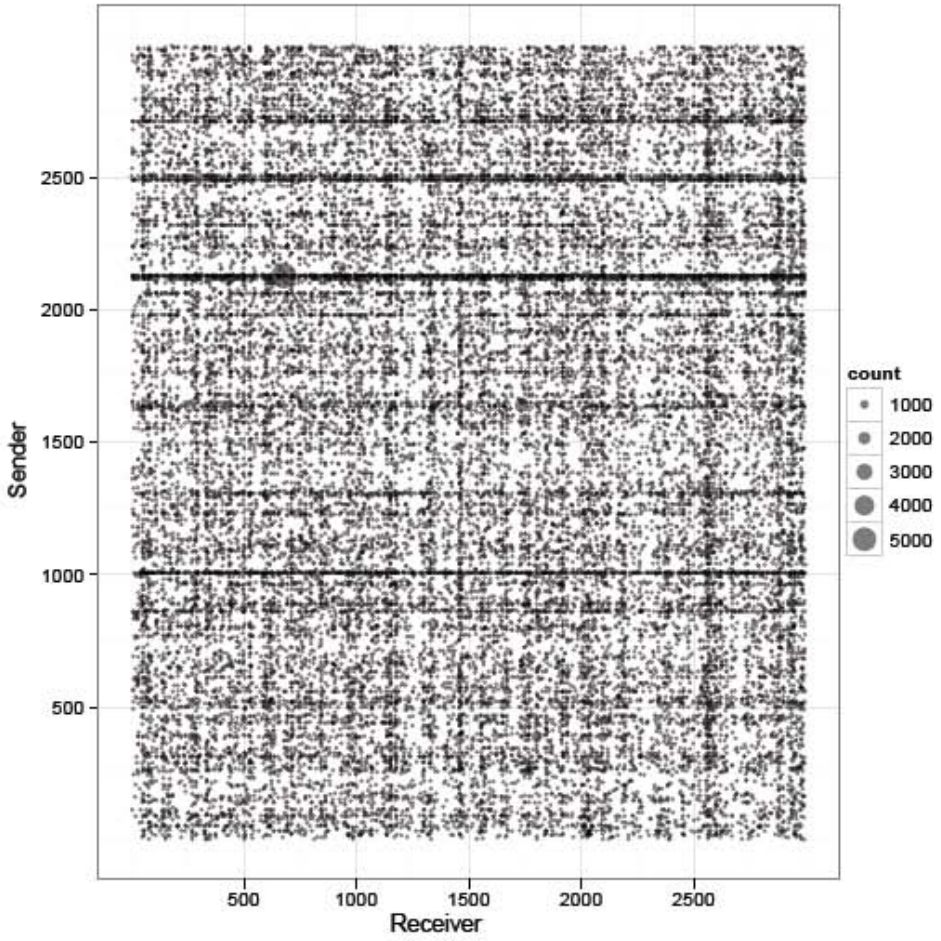


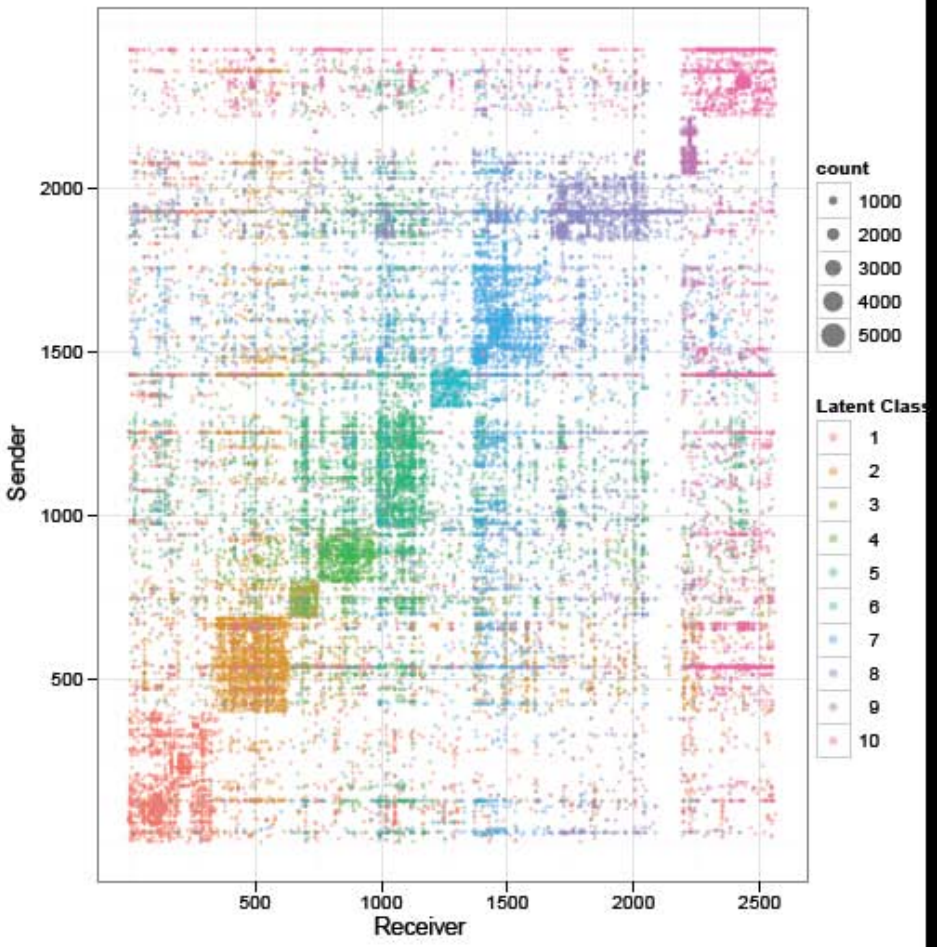
Accomplishments: Scalability

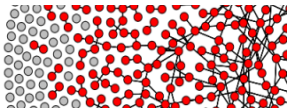
Topic	State of the Art in 2008	State of the Art now (with MURI)	Potential Applications And Impact
Mixtures of ERG models	600 nodes binary-valued (Daudin et al, 2008)	100,000 nodes Categorical-valued (Hunter and Vu, 2010)	Broadens applicability of statistical inference to large noisy networks
Latent variable network models	100 nodes (Raftery et al, JRSS, 2006)	100,000 nodes Efficient latent-class algorithm (DuBois and Smyth, 2010)	Extends statistical network models to data sets where only descriptive methods were used previously

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

Poster by PhD student Chris DuBois







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……
  - More details in Dave Hunter's talk before lunch today

Impact: Publications, Workshops, Talks

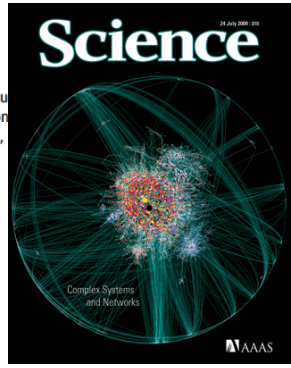
- Over 30 peer-reviewed publications, across computer science, statistics, and social science
  - High visibility
    - *Science,* Butts, 2009
    - *Annals of Applied Statistics*, Gile and Handcock, 2010
    - *Journal of the ACM*, da Fonseca and Mount, 2010
    - *Journal of Machine Learning Research*, 2010
  - Highly selective conferences
    - ACM SIGKDD 2010 (16% accept rate)
    - Neural Information Processing (NIPS) Conference 2009 (25% accepts)
    - IEEE Infocom 2010 (17.5% accepts)

PERSPECTIVE

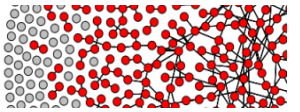
Revisiting the Foundations of Network Analysis

Carter T. Butts

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



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

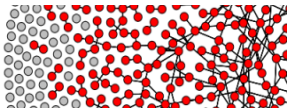


Impact: Workshops and Invited Talks

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

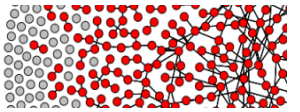
- Invited Talks: Universities
  - Stanford, UCLA, Georgia Tech, U Mass, Brown, etc

- 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…..



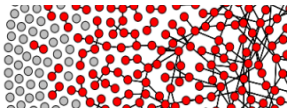
Impact: the Next Generation

- Faculty position at U Mass
  - Ryan Acton -> Asst Prof, part of new initiative in Computational Social Science
- Students speaking at summer 2010 conferences
  - Sunbelt International Social Networks (Jasny, Spiro, Fitzhugh, Almquist, DuBois
  - ACM SIGKDD Conference (DuBois)
  - American Sociological Association Meeting (Marcum, Jasny, Spiro, Fitzhugh, Almquist)
- 2010 Summer school on social network analysis
  - DuBois and Almquist received scholarships to attend
- Best paper awards or nominations (Spiro, Hummel)
- National fellowships (DuBois, Asuncion)



…..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

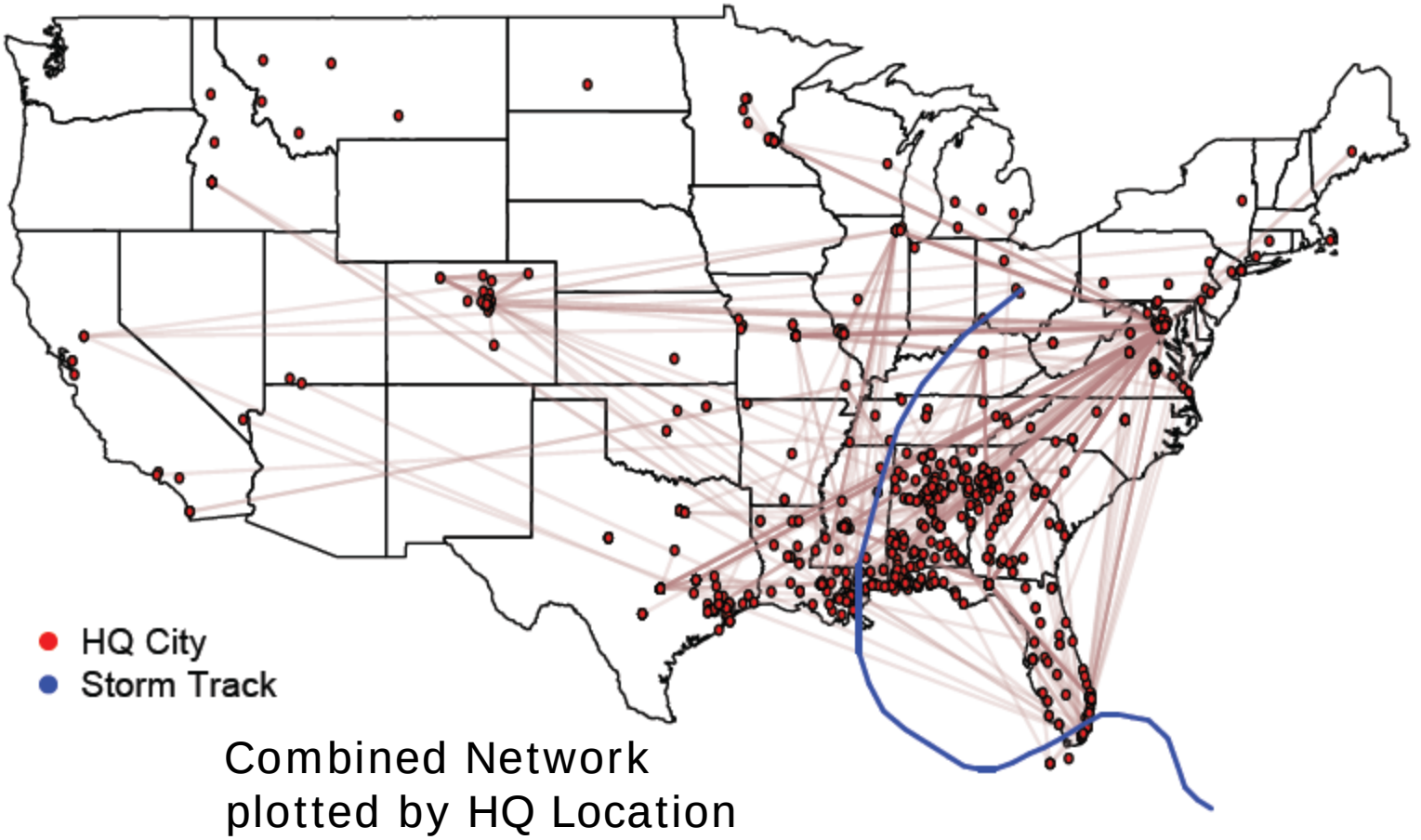


What Next?

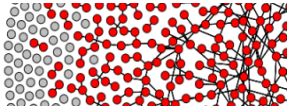
- "Push" algorithm advances into statistical modeling
  - Will allow us to scale existing algorithms to much larger data sets
- Develop network models with richer representational power
  - Geographic data, temporal events, text data, actor covariates, heterogeneity, etc
- Systematically evaluate and 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

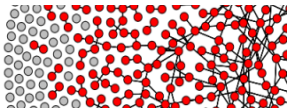
Poster by PhD student Zack Almquist



Combined Network
plotted by HQ Location



Schedule for Today



SESSION 1: SCALABLE METHODS FOR NETWORK MODELING

8:55 Algorithms and Data Structures for Fast Computations on Networks
Mike Goodrich, Professor, Department of Computer Science, UC Irvine

9:20 Listing All Maximal Cliques in Sparse Graphs in Near-Optimal Time
Darren Strash, PhD student, Department of Computer Science, UC Irvine

9:45 Fast Variational Algorithms for Statistical Network Modeling
David Hunter, Professor, Department of Statistics, Penn State University

10:10 Coffee Break

SESSION 2: MODELING SPATIAL, DYNAMIC, AND GROUP STRUCTURE IN NETWORKS

10:30 Efficient Algorithms for Latent Space Embedding
David Mount, Professor, Department of Computer Science, University of Maryland

10:55 Inferring Groups from Communication Data
Chris DuBois, PhD student, Department of Statistics, UC Irvine

11:15 Extended Structures of Mediation: Re-examining Brokerage in Dynamic Networks
Emma Spiro, PhD student, Department of Sociology, UC Irvine

11:35 Update on Publicly Available Software and Data Sets
David Hunter plus graduate students

12:15 Lunch: PIs + visitors at the University Club, Students and Postdocs in 6011

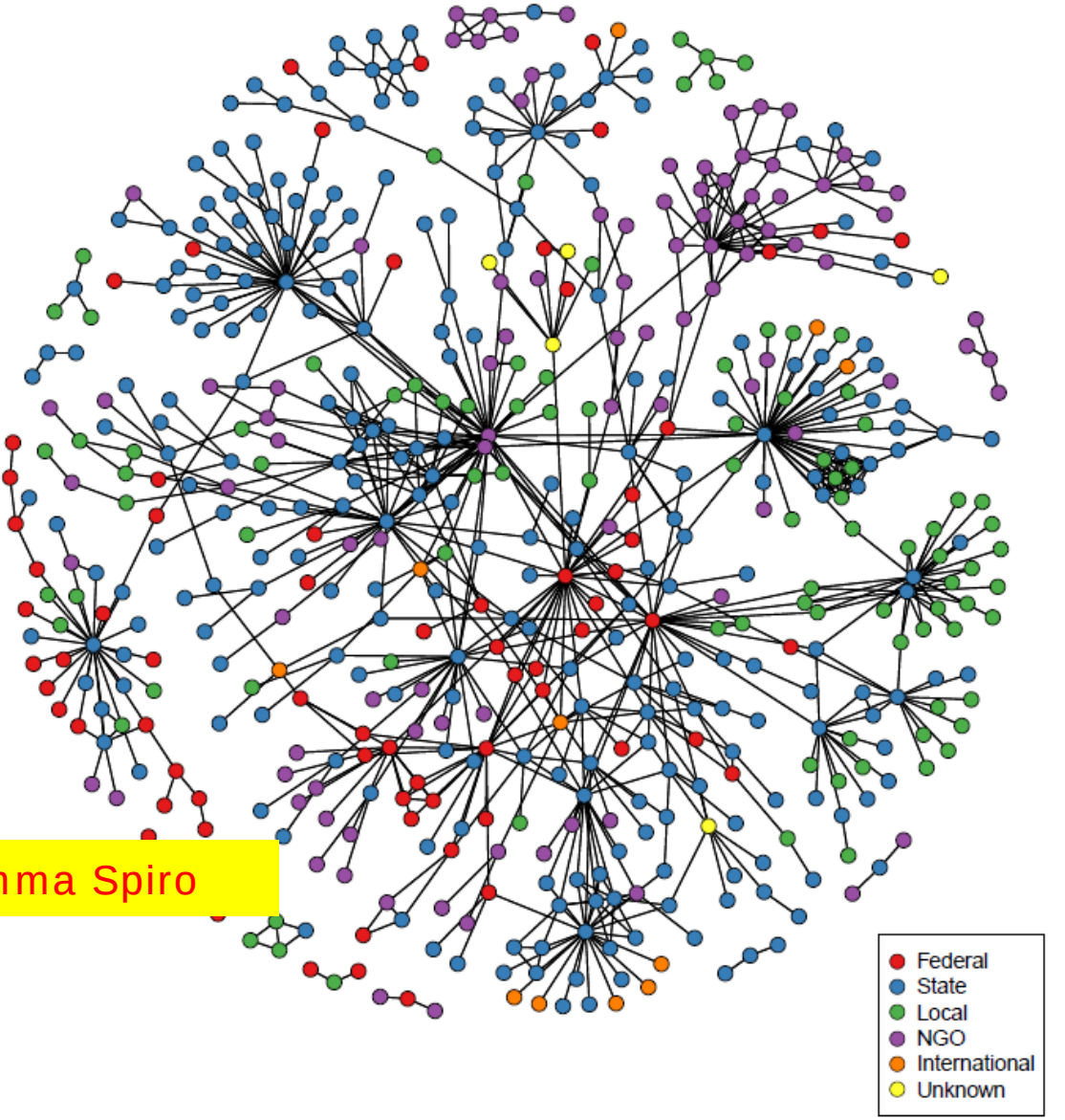
1:15 to 2:45 SESSION 3: POSTERS (see list on next page)

2:45 Advances in Scalable Modeling of Complex, Dynamic Networks
Carter Butts, Professor, Department of Sociology, UC Irvine

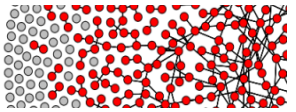
3:10 DISCUSSION AND FEEDBACK

3:30 ADJOURN

Brokerage in Dynamic Networks

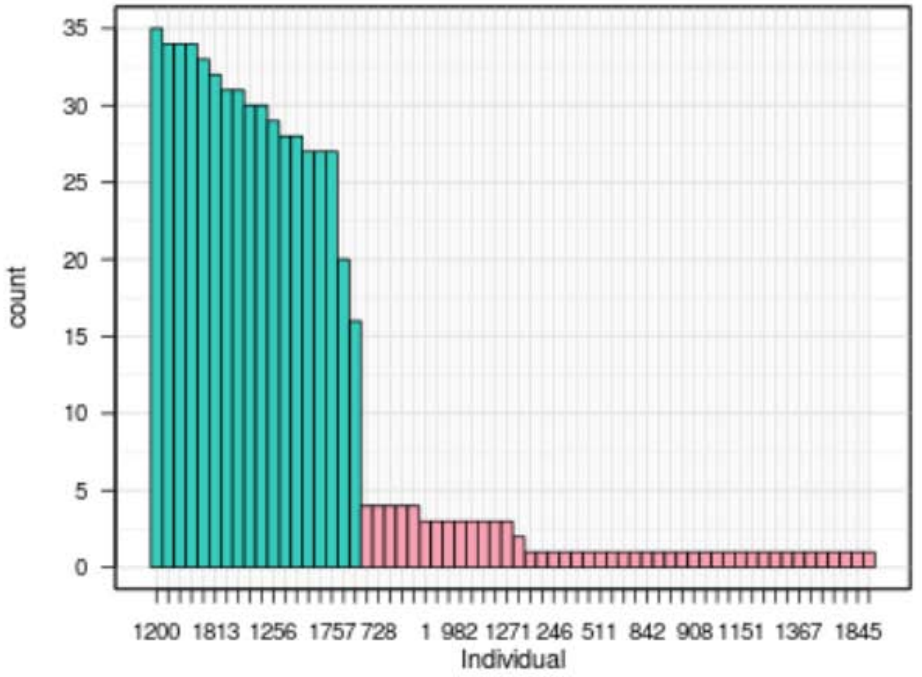


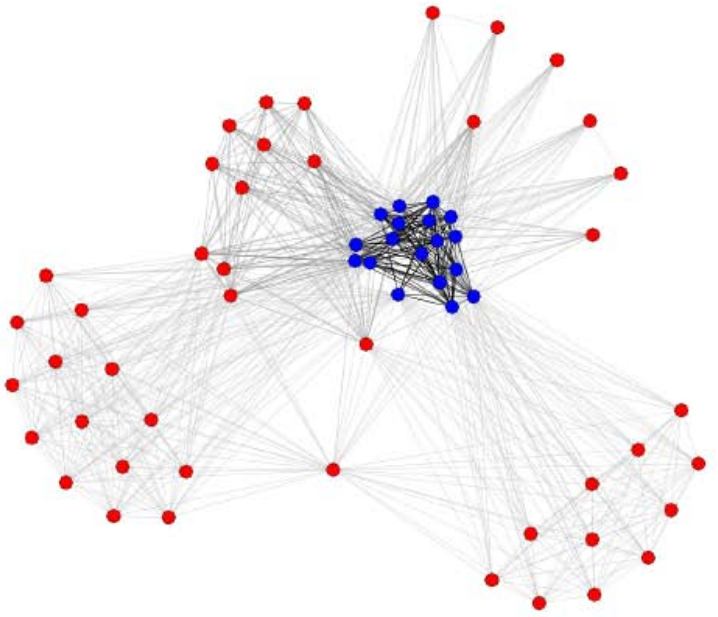
Talk by PhD student Emma Spiro

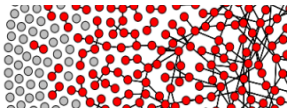


Modeling Groups in Email Communications

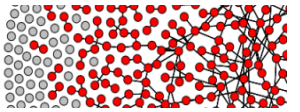
Talk by PhD student Chris DuBois





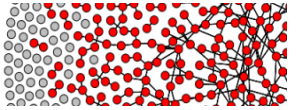


Title	Presenter	Affiliation	Status
Permutation tests for two-mode data	Lorien Jasny	UC Irvine	PhD student
Seasonal modeling of association patterns from time-use data	Chris Marcum	UC Irvine	PhD student
Logistic network regression for scalable analysis of dynamic relational data	Zack Almquist	UC Irvine	PhD student
A network approach to pattern discovery in spell data	Sean Fitzhugh	UC Irvine	PhD student
Rumoring in informal online communication networks	Emma Spiro	UC Irvine	PhD student
Listing all maximal cliques in sparse graphs in near-optimal time	Darren Strash	UC Irvine	PhD student
Extended dynamic subgraph statistics using the h-Index	Lowell Trott	UC Irvine	PhD student
Modeling relational events via latent classes	Chris DuBois	UC Irvine	PhD student
Self-adjusting geometric structures for latent space embedding	Eunhui Park	U Maryland	PhD student
Latent variable models for network data over time	Jimmy Foulds	UC Irvine	PhD student
Hierarchical analysis of relational event data	Nicole Pierski	UC Irvine	PhD student
Retroactive data structures	Joe Simons	UC Irvine	PhD student
Imputing missing data in sensor networks via Markov random fields	Scott Triglia	UC Irvine	PhD student
	Nicholas Navaroli	UC Irvine	PhD student
Viable and non-viable models of large networks, simulation and inference	Michael Schweinberger	Penn State	Postdoctoral Fellow
Bayesian inference and model selection for exponential-family social network models	Ranran Wang	U Washington	PhD student



Logistics

- Meals
  - Lunch at University Club at 12:15 - for visitors and PIs
  - Refreshments at 10:10 and at 2:45
- Wireless
  - Should be able to get 24-hour guest access from UCI network
- Online Slides and Schedule
  www.datalab.uci.edu/muri
  (also contains information about project publications, data sets, software, etc)
- Reminder to speakers: leave time for questions and discussion!



QUESTIONS?