Scalable Methods for the Analysis of Network-Based Data

MURI Project: University of California, Irvine

Project Meeting

April 24th 2009

Principal Investigator: Padhraic Smyth
Goals for Today’s Meeting

• Introductions and brief review of our project

• Technical presentations and discussion
  – MURI-related research
  – Talks from all research groups
  – Long and short talks
  – Important to leave time for questions and discussion
  – Goal is to spur discussion and interaction

• End of day
  – Open discussion on research and collaboration
  – Organizational items
  – 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
PI Collaboration Network

[Diagram showing relationships between individuals: Mike Goodrich, David Eppstein, Padhraic Smyth, Dave Mount, Carter Butts, Dave Hunter, Mark Handcock]
Our Collaboration Network
Statistical Modeling of Network Data

• Statistics = principled approach for inference from data
  – Basis for optimal prediction
    • querying = computation of conditional probabilities/expectation
  – Principles for handling noisy measurements
    • e.g., noisy edge observation process
  – Integration of different sources of information
    • e.g., combining edge information with node covariates
  – Quantification of uncertainty
    • e.g., which model is a better explanation of the data
Limitations of Existing Methods

• Network data over time
  – Relatively few statistical models for 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 model (ERGM)

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

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
Key Themes of our MURI Project

• Foundational research on new statistical estimation techniques for network data
  – e.g., principles of modeling with missing data

• New algorithms for heterogeneous network data
  – Incorporating time, space, text, other covariates

• Faster algorithms
  – E.g., approximation methods for very large data sets

• Software
  – Make network inference software publicly-available (in R)
Key Themes of our MURI Project

- Fast Algorithms
- Statistical Methods
- Richer models
- Large Heterogeneous Data Sets
- Software
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 handling 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
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 a corporate research lab
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 earlier tasks
The ONR Viewpoint

• At our last meeting ONR shared some of their ideas and interests in the area of social network modeling

• Their current applications typically involve characterization of networks via traditional SNA measures, such as betweenness, closeness, etc

• Statistical models such as ERGMs are on their radar but not being currently used

• Significant emphasis on observing networks of individuals over time
ONR Questions

- How does one select the features in an ERGM?
- 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?
Funding Status

• 3-year project with possible extension to 5 years
  • Official start/end dates: May 1 2008 to 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
  – According to ONR:
    • Next installment will not be sent until prior installments are (mostly) spent
  – Plan to spend current funding by March 2010
  – Anticipate next installment will arrive in early 2010
    • But need to show that we are on track to be “spent out” by March
Logistics

• Meals
  – Lunch at University Club - for students, postdocs, faculty
  – Coffee breaks at 10:30 and 3:30
  – Dinner and get-together at Steelhead restaurant, around 5:30, 6:00

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

• Online Slides and Schedule
  http://sites.google.com/site/scalablenetworkanalysis/all-hands-meeting---april-2009
  (Log in to Google sites with your Google password)

• Reminder to speakers: leave time for questions and discussion!