Scalable Methods for the Analysis of Network-Based Data

MURI Project: University of California, Irvine

Project Meeting

May 25th 2010

Principal Investigator: Padhraic Smyth
Today’s Meeting

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

• Format
  – Introduction
  – Research talks
    • 20 and 30 minute slots
    • 5 mins at end for questions/discussion
  – Question/discussion encouraged during talks
  – Several breaks for discussion
Project TimeLine

• Project start/end
  – Start date: May 1 2008
  – End date: April 30 2011/2013

• Meetings
  – Nov 2008: All-Hands Kickoff Meeting
  – April 2009: Working Meeting
  – August 2009: Working Meeting
  – December 2009: All-Hands Annual Review
  – May 2010: Working Meeting
MURI Investigators

Padhraic Smyth UCI
David Eppstein UCI
Carter Butts UCI
Michael Goodrich UCI

Mark Handcock UCLA
Dave Mount U Maryland
Dave Hunter Penn State
Graduate Student Progress

Highlights

- Presenting talks at multiple international conferences this summer
  - Sunbelt International Social Networks conference (Jasny, Spiro, Fitzhugh, Almquist)
  - ACM SIGKDD Conference (DuBois)
  - American Sociological Meeting (Marcum, Jasny, Spiro, Fitzhugh, Almquist)
  - + more

- Workshop organization/instruction
  - Political Networks Conference (Spiro, Fitzhugh, Almquist)

- Summer school on social network analysis
  - DuBois and Almquist received scholarships to attend

- Faculty position at U Mass Amherst (Acton)
- Best paper awards or nominations (Spiro, Hummel)
- National fellowships (DuBois, Asuncion)
Publications

**Fundamentals of Exponential Random Graph Models and Network Analysis**

**Scalable Algorithms for Statistical Network Modeling**


Publications

**Geometric and Spatial Embedding Methods**


Publications

**Dynamic and Relational Event Models**


**Statistical Modeling of Text and Networks**


**Measurement of Large Scale Networks**
Data

Statistical Models

Scalable Algorithms

Evaluation

Software and Applications

P. Smyth: Networks MURI Project Meeting, May 25 2010: 10
Statistical Modeling of Network Data

Statistics = principled approach for inference from noisy data

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

Basis for optimal prediction
- computation of conditional probabilities/expectation

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

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

• 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 $n$
  – Limits practical network sizes to order of $n = 100$ nodes
Computational Efficiency

- Parameter estimation can scale from $O(ne)$ to $O(2^{n(n-1)})$

- Algorithms and data structures for efficient computation
  - H-index for change-score statistics
  - Nets and net-trees
  - Efficient clique-finding algorithms
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., $n = 50$, we have $2^{2450} \sim 10^{245}$ graphs to sum over
Key Themes of our MURI Project

• Research on new statistical estimation techniques and models
  – e.g., principles of modeling and predicting networks over time

• Faster algorithms
  – e.g., efficient data structures and algorithms for very large data sets

• New algorithms for heterogeneous network data
  – Incorporating spatial information, text, other covariates

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

Efficient Algorithms

New Statistical Methods

Richer models

Large Heterogeneous Data Sets

New Applications

Software
Complexities of Real Network Data

• Data types
  – Actors and ties
  – Covariates
  – Temporal events
  – Spatial
  – Text

• Structure
  – Hierarchies and clusters

• Measurement issues
  – Sampling
  – Missing data
DuBois and Smyth, 2010
Enron Email Data

- Messages per week (total)
- Number of senders

Sept 2001 (scandal revealed) to Dec 2001 (bankruptcy)
Daily and weekly variation

- Number of emails
- Time of day

Graphs illustrating daily and weekly variation.
Spatially-Embedded Network Data
Butts, Acton, Almquist, 2009
### Missing Data

Missing Data

Handcock and Gile, 2008

\[
\begin{array}{cccc}
  & A & B & C & D \\
  A & - & 1 & 0 & 0 \\
  B & 0 & - & 1 & 1 \\
  C & 0 & 0 & - & 0 \\
  D & 1 & 1 & 1 & - \\
\end{array}
\]

\[
\begin{array}{cccc}
  & A & B & C & D \\
  A & - & \? & \? & \? \\
  B & \? & - & \? & \? \\
  C & 0 & 0 & - & 0 \\
  D & 1 & 1 & 1 & - \\
\end{array}
\]
Statistical Modeling Frameworks

• Exponential random graph models

• Latent-space models

• Relational event models

All 3 frameworks are related – many talks today will touch on at least one of these frameworks
h-index Data Structures

Eppstein and Spiro, 2009

h-index = maximum number such that
h vertices each have at least h neighbors
h-index Data Structures

Eppstein and Spiro, 2009

h-index = maximum number such that h vertices each have at least h neighbors

H = set of h high-degree vertices
L = remaining vertices

Can use H/L partitioning to efficiently compute and track graph statistics in statistical estimation algorithms
Nets and Net Trees

Cho, Mount, Park, 2009
Fast Sampling Methods

Asuncion et al, 2009
Evaluation and Prediction

- Evaluate algorithms on large real-world data sets
  - Disaster response
    - Katrina communication networks, World Trade Center disaster response data
  - Networks of documents
    - Political blogs, Wikipedia
  - Social activities on the Web
    - Twitter data, Facebook networks, email communication networks
  - International relations
  - ... and more

- Evaluation metrics
  - Computational efficiency
  - Goodness of fit and predictive accuracy
ONR Interests
(adapted from presentation/discussion in Nov 2008 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?
Morning Session I

9:00 Introduction and review of project progress
    Padhraic Smyth (UCI)

9:20 Implementation issues for latent-space embeddings
    David Mount (U Maryland)

9:40 Near-optimal fixed parameter tractability of the Bron-Kerbosch
    algorithm for maximal cliques
    Darren Strash (UCI)

10:10 Methods for analysis of behavioral time-use data
    Chris Marcum (UCI)

10:30 BREAK
Morning Session II

10:50  Mixture models for event-based network data
       Chris DuBois (UCI)

11:10  Static and dynamic robustness in emergency-phase communication networks
       Sean Fitzhugh (UCI)

11:30  Bernoulli graph bounds for general random graph models
       Carter Butts (UCI)

LUNCH BREAK

12:00  Lunch for ALL meeting participants in 4011
Afternoon Session I

1:30 Social network analysis of Twitter data
Emma Spiro (UCI)

2:00 Logistic network regression for scalable analysis of dynamic relational data: an overview and case study
Zack Almquist (UCI)

2:20 Latent feature models for network data over time
Jimmy Foulds (UCI)

2:40 New directions in greedy routing on social networks: the membership dimension
Lowell Trott (UCI)

3:10 BREAK
Afternoon Session II

3:30  Bias-adjusted maximum likelihood estimation methods
      Dave Hunter (Penn State)

3:50  Composite likelihood methods for network estimation
      Arthur Asuncion (UCI)

4:10  Discussion and Wrap-up
      - AHM meeting in November/December
      - collaborative activities
      - action items

4:30  ADJOURN
Logistics

• Meals
  – Lunch in this room, 12 noon
  – Refreshment breaks at 10:30 and 3:10

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

• Slides will be posted online on the project Web site
  www.datalab.uci.edu/muri

• Questions and discussion are encouraged during talks!
Questions?
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


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
Estimation Algorithms

• We want $P(\text{parameters} \mid \text{data})$

• Exact algorithms are rare

• Approximate search
  – E.g., Markov chain Monte Carlo

• Exact solution of simpler objective function
  – E.g., pseudolikelihood v. likelihood
Collaboration Network

- Mike Goodrich
- David Eppstein
- Padhraic Smyth
- Dave Mount
- Carter Butts
- Dave Hunter
- Mark Handcock

P. Smyth: Networks MURI Project Meeting, May 25 2010: 41