Empirical Analysis of Latent Space Embedding

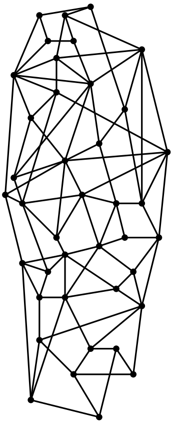
David Mount and Eunhui Park

Department of Computer Science
University of Maryland, College Park

MURI Meeting – June 3, 2011

Motivation

- The likelihood of a tie in social network is often correlated with the similarity of attributes of the actors. (E.g., geography, age, ethnicity, income).
- Attributes may be observed or unobserved (latent).
- We seek to uncover these attributes through the analysis of network's structure.



LSE — Stochastic Model

Input

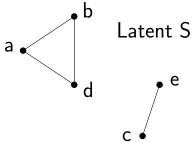
- Y: An $n \times n$ sociomatrix ($y_{i,j} = 1$ if there is a tie between i and j)

Model Parameters

- Z: The positions of n individuals, $\{z_1, \ldots, z_n\}$ in latent space
- α: Real-valued scaling parameter

Network

	a	b	c	d	e
a	-	1	0	1	0
b	1	-	0	1	0
c	0	0	-	0	1
d	1	1	0	-	0
e	0	0	1	0	-



Latent Space

LSE — Stochastic Model

Logistic Regression Model [HRH02]

Ties are statistically independent, and the odds of a tie decreases exponentially with attribute distance.

$$\Pr[Y \mid Z, \alpha] = \prod_{i \neq j} \Pr[y_{i,j} \mid z_i, z_j, \alpha]$$

$$\log \mathrm{odds}(y_{i,j} = 1 \mid z_i, z_j, \alpha) = \alpha - \|z_i - z_j\|.$$

Defining $\eta_{i,j} = \alpha - \|z_i - z_j\|$, we have

$$\log \Pr[Y \mid \eta] = \sum_{i \neq j} (\eta_{i,j} y_{i,j} - \log(1 + e^{\eta_{i,j}})).$$

Optimization

Physical Analogy

Minimize the energy function:

$$-\log \Pr[Y \mid \alpha, \eta] = -\sum_{i \neq j} (\eta_{i,j} y_{i,j} - \log(1 + e^{\eta_{i,j}})),$$

where $\eta_{i,j} = \alpha - \|z_i - z_j\|$.

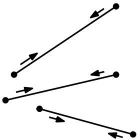
Attractive Component:

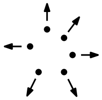
$\sum_{i \neq j} \eta_{i,j} y_{i,j} \Rightarrow$ Avoid long edges

Repulsive Component:

$-\sum_{i \neq j} \log(1 + e^{\eta_{i,j}}) \Rightarrow$ Encourage dispersion

L


Attractive force


Repulsive force

Objective: Find α and $\{z_i\}_{i=1}^{n}$ to minimize energy.
Difficulty: High dimensional and nonlinear.

Approaches

Local Approaches

- Newton-Raphson and gradient descent [NW99]
- Force-directed graph embeddings [BGETT99, B01, FR91]
- Graph layout software [GGK04, GK02, QE01]

Global Approaches

- MCMC-based approaches, like Metropolis-Hastings [HRH02] and simulated annealing

Force-Directed Embedding

Force-Directed Embedding

- for each $u \in V$ do
  - vector $f \leftarrow 0$
  - for each $v \in adj(u)$ do
    - compute attractive strength s_a for edge (u, v)
    - $f \leftarrow f + s_a \cdot \widehat{uv}$
  - for each $v \in V \setminus \{u\}$ do
    - compute repulsive strength s_r for pair $\{u, v\}$
    - $f \leftarrow f + s_r \cdot \widehat{vu}$
  - $pos[u] = pos[u] + f$

where $\widehat{uv}$ is the unit length vector from u to v

Good news: Easy to implement. Tends to converge rapidly

Bad news: Can get stuck in local energy minima

MCMC Algorithm

Markov-Chain Monte-Carlo (MCMC)

- For $k = 0, 1, 2, \ldots$
  - Perturbation: Sample a random perturbation Z_* of Z_k.
  - Evaluation: Compute the decision variable

$$\rho = \frac{\Pr[Y \mid Z_*, \alpha]}{\Pr[Y \mid Z_k, \alpha]}$$

  - Decision: Accept Z_* as Z_{k+1} with probability $\min(1, \rho)$

Good news: Not just a single answer, but provides a sampling of the space of embeddings
Bad news: Hard to know whether you have run long enough to be well mixed

Efficient LSE Computations

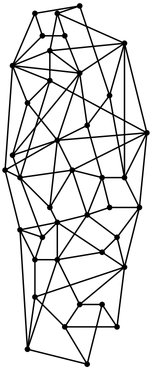
Questions

- What is the nature of local minima?
- How to compute and update forces and change scores efficiently?
- Can we efficiently approximate change scores without adversely affecting MCMC?

- Computation involves retrieval of spatial relations and distances.
- Need efficient geometric retrieval data structures:
  - Approximate: Exact structures are too slow.
  - Incremental: MCMC and force-directed algorithms involve repeated perturbation of point positions.
  - Adaptable: Queries are highly non-uniform, and structures should adapt to these patterns.
  - Variationally Sensitive: Approximations must preserve small variations.

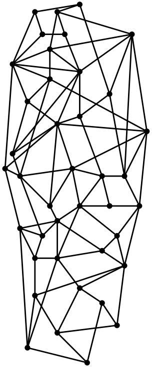
Latent-Space Embedding Exploration Tool

- Our initial attempts provided some successes, some disappointments, and many surprises.
- We needed a better understanding of many issues.
  - What is the nature of the objective function for the logistic model?
  - What sorts of graphs and graph substructures are easy/hard to embed?
  - How robust are embeddings to approximation errors in computing scores?
  - When do force-based algorithms get stuck in local minima and how to extricate them?



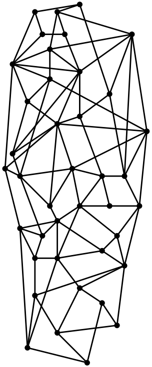
Latent-Space Embedding Exploration Tool

- Our initial attempts provided some successes, some disappointments, and many surprises.
- We needed a better understanding of many issues.
    - What is the nature of the objective function for the logistic model?
    - What sorts of graphs and graph substructures are easy/hard to embed?
    - How robust are embeddings to approximation errors in computing scores?
    - When do force-based algorithms get stuck in local minima and how to extricate them?



Latent-Space Embedding Exploration Tool

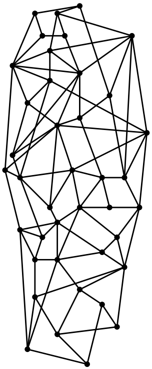
- We are developing an interactive graphical software tool to help us understand, visualize, and experiment with latent-space embeddings
- Similar to the GRIP system of Gajer, Goodrich, and Kobourov [GGK04, GK02]
- Current features:
  - A number of synthetic graph generators (random ala Erdös-Rényi, mesh, torus, logistic-model)
  - A number of force-directed layout algorithms (Fruchterman-Reingold, Hooke's spring law, Eades, logistic-model + gradient descent)
  - User can interactively move and perturb subsets of vertices
  - User can select from various options and parameters



Demo

Latent-Space Embedding Exploration Tool

- Plans:
  - Add MCMC algorithm
  - Provide more graphical instrumentation to determine the algorithm's efficiency and convergence speed
  - Experiment with the effects of variations to algorithm/model/graph parameters



Thank you!

Bibliography

- [BGETT99] G. di Battista, P. Eades, R. Tamassia, I. G. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, 1999.
- [B01] U. Brandes. Drawing on Physical Analogies. In *Drawing Graphs: Methods and Models*. M. Kaufmann and D. Wagner (Eds.), LNCS Tutorial 2025, 71–86. Springer-Verlag, 2001.
- [CK95] P. B. Callahan and S. R. Kosaraju. A decomposition of multidimensional point sets with applications to k-nearest-neighbors and n-body potential fields. *J. Assoc. Comput. Mach.*, 42:67–90, 1995.
- [CMP09] M. Cho, D. M. Mount, and E. Park. Maintaining Nets and Net Trees under Incremental Motion. ISAAC'09, Springer Lecture Notes LNCS 5878, 1134-1143.
- [FR91] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software Practice & Experience* 21: 1129-1164, 1991.
- [GGK04] P. Gajer, M. T. Goodrich, and S. G. Kobourov. A Multi-Dimensional Approach to Force-Directed Layouts of Large Graphs. *CGTA*, 29, 3–18, 2004.

Bibliography

- [GK02] P. Gajer and S. G. Kobourov. GRIP: Graph Drawing with Intelligent Placement *Journal of Graph Algorithms and Applications* 6, 203–224, 2002.
- [HRH02] P. D. Hoff, A. E. Raftery, and M. S Handcock. Latent space approaches to social network analysis. *J. American Statistical Assoc.*, 97:1090–1098, 2002.
- [HRT07] M. S. Handcock and A. E. Raftery and J. M. Tantrum. Model-based clustering for social networks. *J. R. Statist. Soc. A*, 170, Part 2, 301–354, 2007.
- [MNP+04] D. M. Mount, N. S. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu. A computational framework for incremental motion. In *Proc. 20th Annu. ACM Sympos. Comput. Geom.*, 200–209, 2004.
- [NW99] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer-Verlag, 1999.
- [QE01] A. Quigley and P. Eades. FADE: Graph Drawing, Clustering, and Visual Abstraction. *Graph Drawing*, LNCS 1984, 77-80, 2001.