Community Structure in Large Social and Information Networks

Michael W. Mahoney

Stanford University

(For more info, see: http://cs.stanford.edu/people/mmahoney)

Lots and lots of large data!

- DNA micro-array data and DNA SNP data
- High energy physics experimental data
- Hyper-spectral medical and astronomical image data
- Term-document data
- Medical literature analysis data
- Collaboration and citation networks
- Internet networks and web graph data
- Advertiser-bidded phrase data
- Static and dynamic social network data

Networks and networked data

Lots of "networked" data!!

- technological networks
 - AS, power-grid, road networks
- biological networks
 - food-web, protein networks
- social networks
 - collaboration networks, friendships
- information networks

- co-citation, blog cross-postings, advertiser-bidded phrase graphs...

language networks

• ...

- semantic networks...

Interaction graph model of networks:

- Nodes represent "entities"
- Edges represent "interaction" between pairs of entities



Algorithmic vs. Statistical Perspectives

Lambert (2000)

Computer Scientists

- Data: are a record of everything that happened.
- Goal: process the data to find interesting patterns and associations.
- Methodology: Develop approximation algorithms under different models of data access since the goal is typically computationally hard.

Statisticians

- Data: are a particular random instantiation of an underlying process describing unobserved patterns in the world.
- Goal: is to extract information about the world from noisy data.
- Methodology: Make inferences (perhaps about unseen events) by positing a model that describes the random variability of the data around the deterministic model.

Perspectives are NOT incompatible

• Statistical/probabilistic ideas are central to recent work on developing improved randomized algorithms for matrix problems.

• Intractable optimization problems on graphs/networks yield to approximation when assumptions are made about network participants.

• In boosting (a statistical technique that fits an additive model by minimizing an objective function with a method such as gradient descent) the computation parameter (i.e., the number of iterations) also serves as a regularization parameter.

Sponsored ("paid") Search

Text-based ads driven by user query

🕲 recipe indian food - Yahoo! Search Results - Mozilla Firefox	_ 2 2 🔀
<u>File E</u> dit <u>V</u> iew Hi <u>s</u> tory <u>B</u> ookmarks <u>Y</u> ahoo! <u>T</u> ools <u>H</u> elp	\sim
<	▶ G • indian food recipes
🖉 Rutgers University Li 🗋 my del.icio.us 🗋 post to del.icio.us	
MN - powered by MICOL SEARCH + Q Web Search - 2 😥	▼ 👼 Storage 👻
Y - 🖉 - recipe indian food - 🔹 🔶 Search Web - 🔶 - 🔯 Mail - 💽 - 🐼 My Yahoo! 🕉 NCAA Hoops - 🦞 Fant	asy Sports 🔻 📥 Games 🔹 🧞 Music 🔹 🛛 🚿
Yahoo! My Yahoo! Mail Welcome, Guest [Sign In]	Advertiser Sign In Help
Web Images Video Local Shopping more Video Search recipe indian food Search	Answers
Search Results 1 - 10 of about 7,260,000 for re	cipe indian food - 0.19 sec. (<u>About this page</u>)
Recipe Indian Food www.MonsterMarketplace.com - Browse and compare great deals on recipe indian food. Indian Food sanfrancisco.citysearch.com - Find great Indian restaurants in your area today. Search here.	SPONSOR RESULTS Indian Food Buy indian food at SHOP.COM. Search our free shipping offers. www.SHOP.com
1. <u>indian food recipe</u> indian food recipe Title: Indian Food Recipe. Yield: 4 Servings. Ingredients. 1 bunch to the echo by: Jonathan Kandell Indian Food Recipes Put recipes.chef2chef.net/recipe-archive/43/231458.shtml - 13k - <u>Cached</u> - <u>More from this site</u>	Recipe India Food Find and Compare prices on recipe india food at Smarter.com. www.smarter.com
 Recipe Gal: Indian Foods Indian Recipes from Recipe Gal's Archives All Food Posters. Travel Posters. Indian Recipes. Indian Breads Indian Chicken Recipes www.recipegal.com/indian - 10k - <u>Cached</u> - <u>More from this site</u> 	Chinese Food Recipe Books on Cataloglink Find chinese food recipe books on CatalogLink. www.CatalogLink.com
 Indian Recipes, Indian Food Recipe, South Indian Recipes, Indian indian recipes, indian food recipe, south indian Recipes, indian cooking Recipes, Indian Recipes, Indian Food Recipe, South Indian Recipes, Indian Cooking Recipe, www.india4world.com/indian-recipe - 17k - <u>Cached</u> - <u>More from this site</u> Paav Bhaaji - Recipe for Paav Bhaaji - Pao Bhaji 	\$19.97 Over 500 Chinese Recipes Cookbook 100% Satisfaction Guaranteed, 543-Page Chinese Cookbook Only \$19.97. ✓

Sponsored Search Problems

Keyword-advertiser graph:

- provide new ads
- maximize CTR, RPS, advertiser ROI

"Community-related" problems:

Marketplace depth broadening:

Keywords Advertisers bids, clicks or impressions www.allbets.com www.soccer.com soccer videos sports movies hollywood hits www.netflix.com

find new advertisers for a particular query/submarket

• Query recommender system:

suggest to advertisers new queries that have high probability of clicks

Contextual query broadening:

broaden the user's query using other context information

Micro-markets in sponsored search

Goal: Find *isolated* markets/clusters with *sufficient money/clicks* with *sufficient coherence*. Ques: Is this even possible?



10 million keywords

What do these networks "look" like?



Questions of interest ...

What are degree distributions, clustering coefficients, diameters, etc.? Heavy-tailed, small-world, expander, geometry+rewiring, local-global decompositions, ... Are there natural clusters, communities, partitions, etc.? Concept-based clusters, link-based clusters, density-based clusters, ... (e.g., isolated micro-markets with sufficient money/clicks with sufficient coherence) How do networks grow, evolve, respond to perturbations, etc.? Preferential attachment, copying, HOT, shrinking diameters, ... How do dynamic processes - search, diffusion, etc. - behave on networks? Decentralized search, undirected diffusion, cascading epidemics, ... How best to do learning, e.g., classification, regression, ranking, etc.? Information retrieval, machine learning, ...

Clustering and Community Finding

• Linear (Low-rank) methods

If Gaussian, then low-rank space is good.

• Kernel (non-linear) methods

If low-dimensional manifold, then kernels are good

Hierarchical methods

Top-down and botton-up -- common in the social sciences

• Graph partitioning methods

Define "edge counting" metric -- conductance, expansion, modularity, etc. -- in interaction graph, then optimize!

"It is a matter of common experience that communities exist in networks ... Although not precisely defined, communities are usually thought of as sets of nodes with better connections amongst its members than with the rest of the world."



Community Score: Conductance

How community like is a set of nodes?

Conductance (normalized cut)

Need a natural intuitive measure:



Community Score: Conductance



Score: $\phi(S) = #$ edges cut / # edges inside



Score: $\phi(S) = #$ edges cut / # edges inside ¹⁴



Score: $\phi(S) = #$ edges cut / # edges inside



Score: $\phi(S) = #$ edges cut / # edges inside

16

Network Community Profile Plot

We define:

Network community profile (NCP) plot

Plot the score of best community of size k

$$\Phi(k) = \min_{S \subset V, |S|=k} \phi(S)$$

- Search over all subsets of size k and find best: (k=5) = 0.25
- NCP plot is intractable to compute
- Use approximation algorithms



Widely-studied small social networks





Newman's Network Science

"Low-dimensional" graphs (and expanders)



What do large networks look like?

Downward sloping NCPP

small social networks (validation)

"low-dimensional" networks (intuition)

hierarchical networks (model building)



Natural interpretation in terms of isoperimetry

implicit in modeling with low-dimensional spaces, manifolds, k-means, etc.

Large social/information networks are very very different

We examined more than 70 large social and information networks We developed principled methods to interrogate large networks Previous community work: on small social networks (hundreds, thousands)

Large Social and Information Networks

• Social nets	Nodes	Edges	Description
LIVEJOURNAL	4,843,953	42,845,684	Blog friendships [4]
Epinions	75,877	405,739	Who-trusts-whom [35]
FLICKR	404,733	2,110,078	Photo sharing [21]
Delicious	147,567	301,921	Collaborative tagging
CA-DBLP	317,080	1,049,866	Co-authorship (CA) [4]
CA-COND-MAT	21,363	91,286	CA cond-mat [25]
• Information networks			
CIT-HEP-TH	27,400	352,021	hep-th citations [13]
Blog-Posts	437,305	565,072	Blog post links [28]
• Web graphs			
Web-google	855,802	4,291,352	Web graph Google
Web-wt10g	1,458,316	6,225,033	TREC WT10G web
• Bipartite affiliation (authors-to-papers) networks			
ATP-DBLP	615,678	944,456	DBLP [25]
ATP-ASTRO-PH	54,498	131,123	Arxiv astro-ph [25]
• Internet networks			
AS	6,474	12,572	Autonomous systems
GNUTELLA	62,561	147,878	P2P network [36]

Table 1: Some of the network datasets we studied.

Approximation algorithms as experimental probes?

The usual modus operandi for approximation algorithms for general problems:

- define an objective, the numerical value of which is intractable to compute
- develop approximation algorithm that returns approximation to that number
- graph achieving the approximation may be unrelated to the graph achieving the exact optimum.

But, for randomized approximation algorithms with a geometric flavor (e.g. matrix algorithms, regression algorithms, eigenvector algorithms; duality algorithms, etc):

- often can approximate the vector achieving the exact solution
- randomized algorithms compute an ensemble of answers -- the details of which depend on choices made by the algorithm
- maybe compare different approximation algorithms for the same problem.

Probing Large Networks with Approximation Algorithms

Idea: Use approximation algorithms for NP-hard graph partitioning problems as experimental probes of network structure.

Spectral - (quadratic approx) - confuses "long paths" with "deep cuts" Multi-commodity flow - (log(n) approx) - difficulty with expanders SDP - (sqrt(log(n)) approx) - best in theory Metis - (multi-resolution for mesh-like graphs) - common in practice X+MQI - post-processing step on, e.g., Spectral of Metis

Metis+MQI - best conductance (empirically)

Local Spectral - connected and tighter sets (empirically, regularized communities!)

We are not interested in partitions per se, but in probing network structure.

Analogy: What does a protein look like?



Three possible representations (all-atom; backbone; and solvent-accessible surface) of the three-dimensional structure of the protein triose phosphate isomerase.

Experimental Procedure:



- Generate a bunch of output data by using the unseen object to filter a known input signal.
- Reconstruct the unseen object given the output signal and what we know about the artifactual properties of the input signal.

Typical example of our findings

General relativity collaboration network (4,158 nodes, 13,422 edges)



Large Social and Information Networks



Focus on the red curves (local spectral algorithm) - blue (Metis+Flow), green (Bag of whiskers), and black (randomly rewired network) for consistency and cross-validation.

More large networks



10⁶

10⁵

NCPP: LiveJournal (N=5M, E=43M)



"Whiskers" and the "core"

- "Whiskers"
 - maximal sub-graph detached from network by removing a single edge
 - contains 40% of nodes and 20% of edges
- "Core"
 - the rest of the graph, i.e., the
 2-edge-connected core
- Global minimum of NCPP is a whisker



What if the "whiskers" are removed?

Then the lowest conductance sets - the "best" communities - are "2-whiskers." (So, the "core" peels apart like an onion.)



Lower Bounds ...

- ... can be computed from:
- Spectral embedding

(independent of balance)

SDP-based methods

(for volume-balanced partitions)





NCPP for common generative models



A simple theorem on random graphs

Let $\mathbf{w} = (w_1, \dots, w_n)$, where $w_i = ci^{-1/(\beta-1)}, \quad \beta \in (2,3).$ Connect nodes *i* and *j* w.p. $p_{ij} = w_i w_j / \sum_k w_k.$





Structure of the G(w) model, with $\beta \epsilon$ (2,3).

- Sparsity (coupled with randomness) is the issue, not heavy-tails.
- (Power laws with $\beta \epsilon$ (2,3) give us the appropriate sparsity.)

A "forest fire" model

Model of: Leskovec, Kleinberg, and Faloutsos 2005

At each time step, iteratively add edges with a "forest fire" burning mechanism.





Also get "densification" and "shrinking diameters" of real graphs with these parameters (Leskovec et al. 05).

Comparison with "Ground truth" (1 of 2)

Networks with "ground truth" communities:

- LiveJournal12:
 - users create and explicitly join on-line groups
- CA-DBLP:
 - publication venues can be viewed as communities
- AmazonAllProd:
 - each item belongs to one or more hierarchically organized categories, as defined by Amazon
- AtM-IMDB:
 - countries of production and languages may be viewed as communities (thus every movie belongs to exactly one community and actors belongs to all communities to which movies in which they appeared belong)





AtM-IMDB

 10^{4}

10⁵

10⁶

10⁵

"Structure" and "randomness" in very large informatics graphs

Mahoney and Leskovec (2009)

Some high-level themes to formalize:

• There do not exist a "small" number of linear components that capture "most" of the variance/information in the data.

• There do not exist "nice" manifolds that describe the data well.

• There is "locally linear" structure or geometry on small size scales that does not propagate to global/large size scales.

• At large size scales, the "true" geometry is more "hyperbolic" or "tree-like" or "expander-like".

Important: even if you do not care about communities, conductance, etc., these empirical observations place very severe constraints on the types of models that are appropriate to consider.

"Learning" with Spectral Kernels and Heavy-Tailed Data

Mahoney and Narayanan (2009)

Recall the usual story:

• The sample complexity for distribution-free learning typically depends on the ambient dimension to which the data to be classified belongs.

• For very high-dimensional data, such bounds can be unsatisfactory, motivating discussions of manifolds, etc.

Motivated by informatics graphs, we have several novel issues:

- Heavy-tailed distributions of degrees, eigenvalues, etc.
- Spectral (especially local spectral) kernels good for denoising, etc.
- Data are naturally modeled by a graph, not a feature vector.

"Learning" with Spectral Kernels and Heavy-Tailed Data

Mahoney and Narayanan (2009)

Heavy-tailed data - e.g., graphs with heavy-tailed degree distributions

• A small number of big guys, but many many little guys at every lower scale, so neither dominates.

Spectral Kernels - e.g., Laplacian Eigenmaps, diffusion maps, etc.

- Construct kernel from eigenvalues/eigenfunctions graph Laplacian.
- Entries of eigenvectors (and thus of the kernel) are NOT uniformly bounded in general since they may be localized.

In both cases:

- Dstbn-indep VC arguments give trivial dimension-dependent results.
- Can use dstbn-dependent annealed entropy to provide dimension-indep learning bounds!

Conclusions

Approximation algorithms as experimental probes!

- Hard-to-cut onion-like core with more structure than random
- Small well-isolated communities gradually blend into the core

Community structure in large networks is qualitatively different!

- Agree with previous results on small networks
- Agree with sociological interpretation (Dunbar's 150 and bond vs. identity)!

Common generative models don't capture community phenomenon!

- Graph locality important for realistic network generation
- Local regularization important due to sparsity