Graphlets

Edo Airoldi

Department of Statistics
Harvard

Algorithms for Massive Data, 13 July 2012, Palo Alto, CA

Overview

- Structured data vs. latent dependence structure
  - Leveraging observed (noisy) structure for estimation
- This talk
  - Structure is expressed by a weighted network
- Challenges and goals
  - Structured (eg, paired) measurements and interference
  - Expressive/interpretable model, scalable inference, theory

Agenda

- Graphlets
  1. Theory
  2. Empirical results
- Concluding remarks

Quantifying social information

- A representation problem
  - How to translate social structure into (social) information?
- Technical challenges for semi-parametric approach
  - Basis must have substantive interpretation to be useful
  - Theoretically very large number of basis elements
- Graphlets
  - Basis encodes notion of multiple/multi-scale membership
  - Induce the set of feasible basis elt's from data

Problem definition

- Consider a network with N nodes, with adjacency matrix D with integer entries.
- The matrix D could results from counting events on pairs of nodes over a fixed period of time
- We want a model of network structure in which
  - Groups are a function of (unobservable) nodes properties
  - Properties may induce groups at multiple scales
  - Edge weights combine multiple/multi-scale groups

Basic Graphlets semi-par. model

- Assume a distribution for the edge-weights

$$\Pr(D \mid B, \mu) = \prod_{ij} \text{Poisson}(\lambda_{ij})$$

$\lambda_{ij} = B'MB$; D is $(N \times N)$; B is $(K \times N)$ and sparse

- Assume an arbitrary distribution on K

Basic Graphlets decomposition

- Intuitively
  - N nodes are represented as a strings of K-bits
  - K groups are induced shared bits
  - Nested (multi-scale) groups and overlapping (multiple membership) groups are induced bit combinations
- Mathematically, a dual representation
  - Node/bit strings: $\Lambda_{ij} = b'M\, b_{(K\times 1)}$ and $\Lambda = B'M\, B_{(K\times N)}$
  - Network/groups: $\Lambda = \sum_{i=1\ldots K} \mu_i \cdot P_i(B)_{(N\times N)}$

Inference 1: almost finding B

- Induce a candidate set of basis elt's from data, B_c

B_c=empty Basis set
Generating B_c**:**
  For(level=$max(D)$ *to* $min(D)$)
    $D_{th=level} = \mathbf{1}(D \geq level)$
    C=maximal cliques($D_{th=level}$) using Bron-Kerbosch algorithm
    B_c=$B_c \cup C$

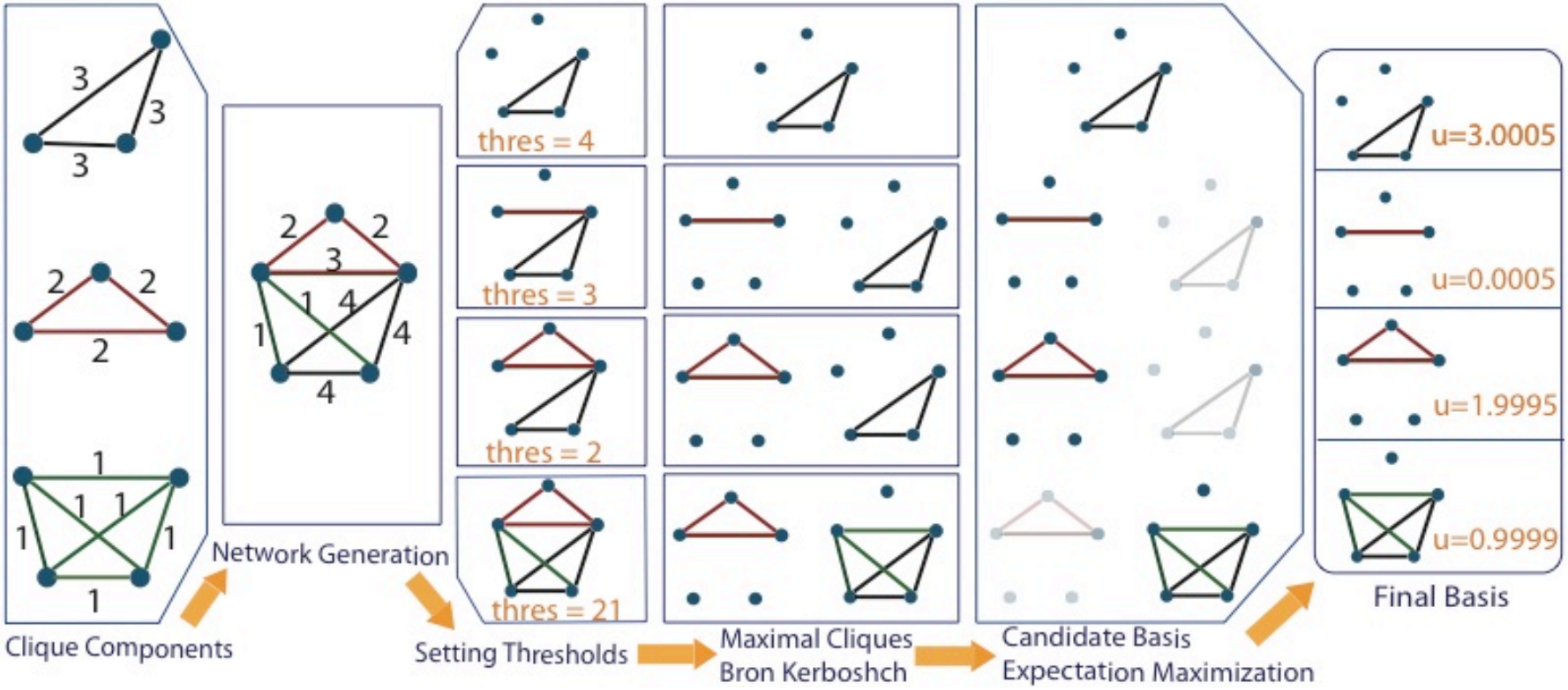
Algorithm 1: Algorithm for generating matrix B_c from the data D

Theorem 1: $B \subseteq B_c$ with high probability – under regularity conditions on B at successive levels

Inference 2: finding the μ's (and B)

- At this stage we have: $D = \sum_{i=1 \ldots Kc} \mu_i \cdot P_i(B_c)$ with more elt's P_i than we'd expect to have
- We develop an Expectation – Maximization (EM) algorithm to estimate all μ_i
- The EM algorithm will zero out redundant μ_i's, effectively estimating B – the P_i's with $\mu_i > 0$

An illustration



Agenda

- Graphlets
  1. Theory
  2. Empirical results
- Concluding remarks

Complexity and redundancy of B_c

Theorem 3. *Let the elements B_{ik} of the basis matrix for a network Y be IID Bernoulli random variables with parameter p_N. Then an asymptotic upper bound for the number of candidate basis elements, denoted C_{N,p_N}, identified by Algorithm 1 is*

$$\begin{aligned} C_{N,p_N} &\leq Q(2^{KH(p_N)} + K) \\ &= Q(N^{c_1 H(p_N)} + c \log_2 N), \end{aligned} \tag{2}$$

where Q is the number of thresholds in Algorithm 1.

- In the case where $p = O(1/\log_2 N)$, the redundancy is $R_N = O(1)$, and the total number of basis elt's is $O(\log N)$ – the same order of magnitude of K
- This seems to hold in practice, since $C_{N,p} \approx 2K$

Scalability

First stage

O (size of the biggest clique × no. of levels)

Second stage

Dimensionality is reduced from no. of positive D weights to the no. of elements in B_c – in practice $C_{N,p} = O(K)$

The EM algorithm converges in a few iterations

Overall inference is *O (sub no. edges)* in practice, and can scale to very large weighted networks

Theoretical accuracy

Theorem 4. *The theoretical accuracy of the best approximate Graphlet decomposition with $\tilde{K}$ out of K basis elements is:*

$$\tau_0(\tilde{K}, K, \alpha) = \sum_{j=1}^{\tilde{K}} \frac{f(j, K, \alpha)}{\alpha K}, \tag{7}$$

where

$$f(j, K, \alpha) = \binom{K}{j} \sum_{q=0}^{j-1} (-1)^q \binom{j-1}{q} \frac{f(1, K-j+q+1, \alpha)}{K-j+q+1} \tag{8}$$

$$f(1, K, \alpha) = \frac{K}{\Gamma(\alpha)} \sum_{m=0}^{(\alpha-1)(K-1)} c_m(\alpha, K-1) \frac{\Gamma(\alpha+m)}{K^{\alpha+m}}, \tag{9}$$

in which the coefficients c_m are defined by the recursion

$$c_m(\alpha, q) = \sum_{i=0}^{i=\alpha-1} \frac{1}{i!} c_{m-i}(\alpha, q-1) \tag{10}$$

with boundary conditions $c_m(\alpha, 1) = \frac{1}{i!}$ for $i = 1 \ldots \alpha$.

Agenda

- Graphlets
  1. Theory
  2. Empirical results
- Concluding remarks

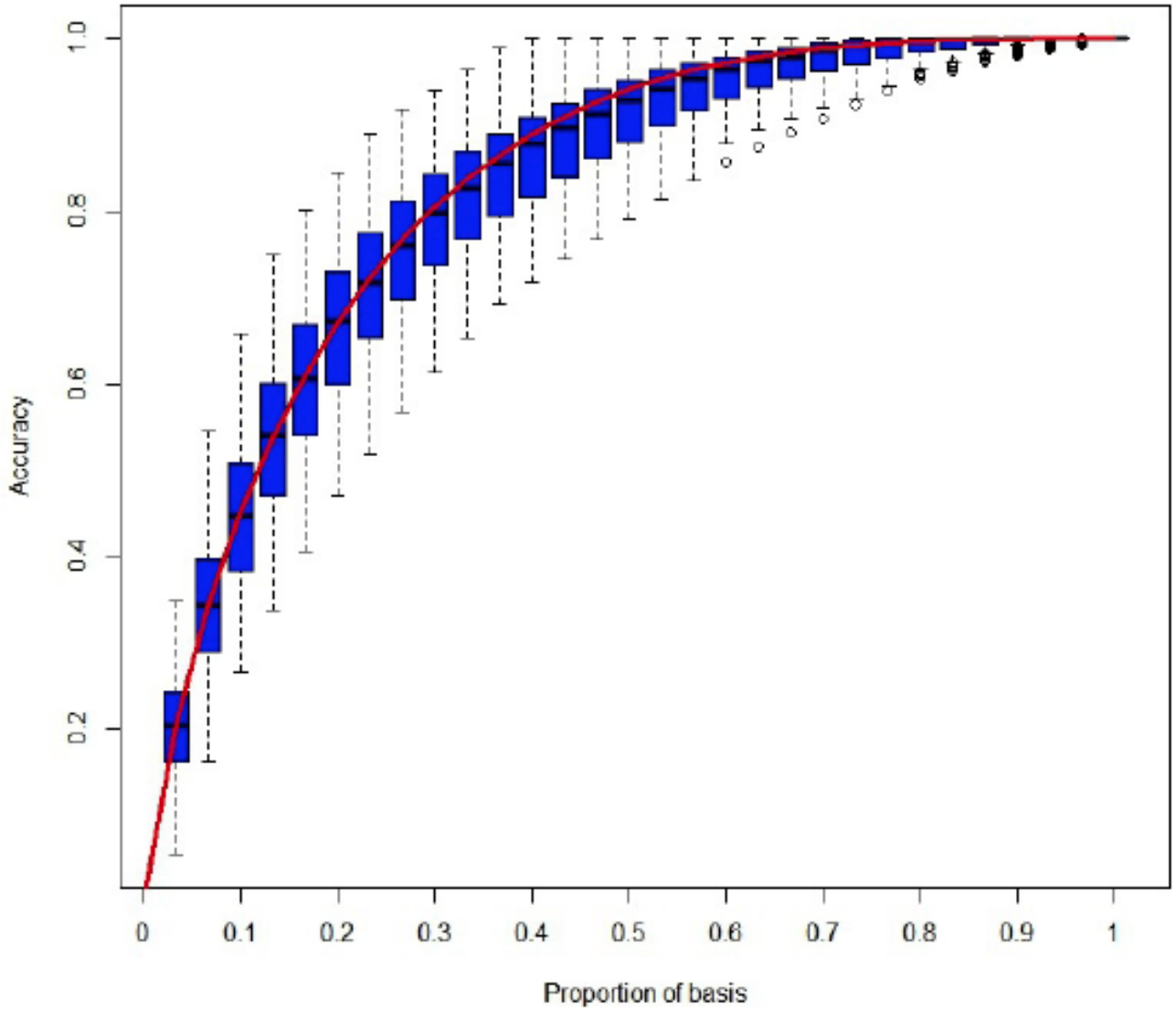


Figure 2: Theoretical and empirical accuracy for different fractions of basis elements $\tilde{K}/K$ with $\alpha = 0.1$. The ratio $\tilde{K}/K$ also provides a measure of sparsity.

Evaluation: link prediction

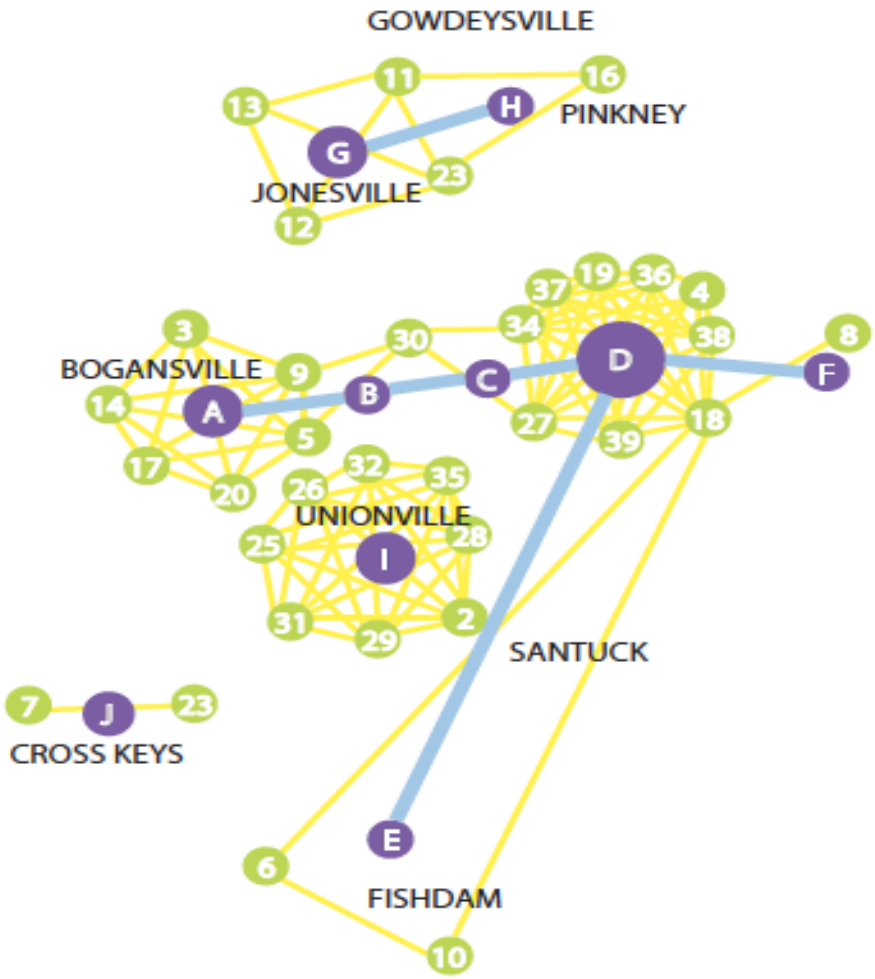
Method	Runtime (sec)	Accuracy
Glet (25%)	0.0636	92.7 ± 0.30
Glet (50%)	0.0636	94.7 ± 0.15
Glet (75%)	0.0636	97.0 ± 0.08
Glet (90%)	0.0636	98.9 ± 0.05
Glet (100%)	0.0636	100.0 ± 0.00
ERG	0.0127	86.0 ± 1.20
DDS	11.1156	89.0 ± 0.85
LSCM	6.3491	90.0 ± 0.70
MMSB	2.8555	93.0 ± 0.40

Analysis of college networks

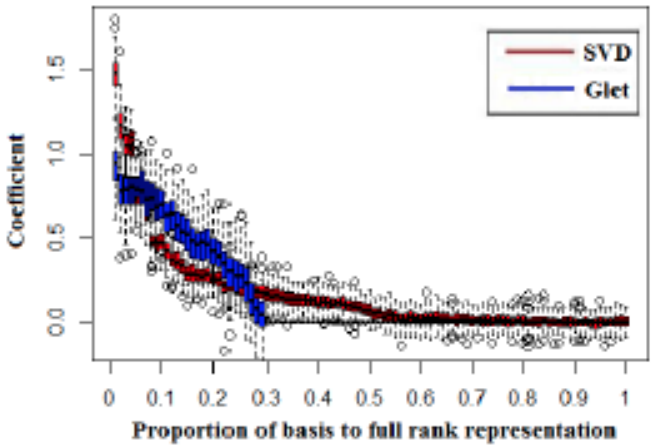
Table 3: $\tau(\tilde{Y})$ error for different fractions of the estimated optimal number of basis elements $\hat{K}$.

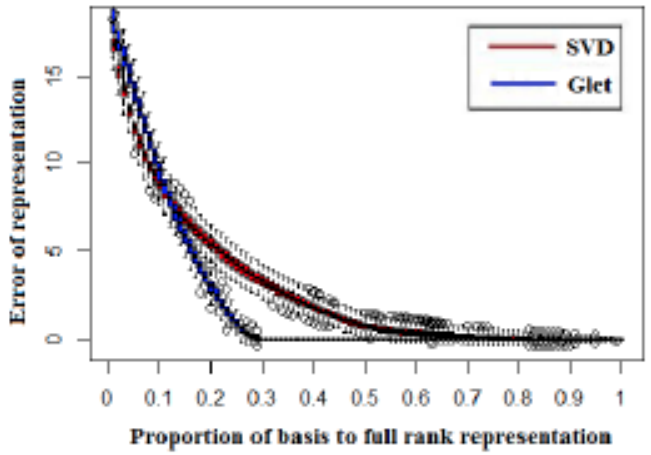
college	nodes	edges	$\hat{K}$	sec	10%	25%	50%	75%	90%	100%
American	6386	435323	11426	151	13.5	6.5	1.80	.70	.30	0
Amherst	2235	181907	10151	124	6.5	2.3	.84	.33	.14	0
Bowdoin	2252	168773	9299	113	9.0	3.2	1.17	.41	.17	0
Brandeis	3898	275133	10340	116	6.8	2.9	1.21	.53	.25	0
Bucknell	3826	317727	13397	193	7.8	2.9	1.15	.45	.20	0
Caltech	769	33311	3735	51	5.7	1.8	.65	.27	.11	0
CMU	6637	499933	11828	169	14.9	5.2	1.89	.73	.34	0
Colgate	3482	310085	12564	151	8.0	3.3	1.26	.45	.19	0
Hamilton	2314	192787	11666	200	6.9	2.5	.84	.33	.15	0
Haverford76	1446	119177	9021	128	4.8	2.2	.74	.25	.10	0
Howard	4047	409699	12773	170	8.6	3.7	1.55	.60	.28	0
Johns Hopkins	5180	373171	11674	150	10.8	3.7	1.40	.58	.29	0
Lehigh	5075	396693	14076	206	9.5	3.2	1.14	.49	.23	0
Michigan	3748	163805	5561	54	11.4	4.6	1.92	.76	.35	0
Middlebury	3075	249219	9971	109	9.7	3.5	1.35	.49	.22	0
MIT	6440	502503	13145	191	11.5	4.7	1.68	.65	.30	0

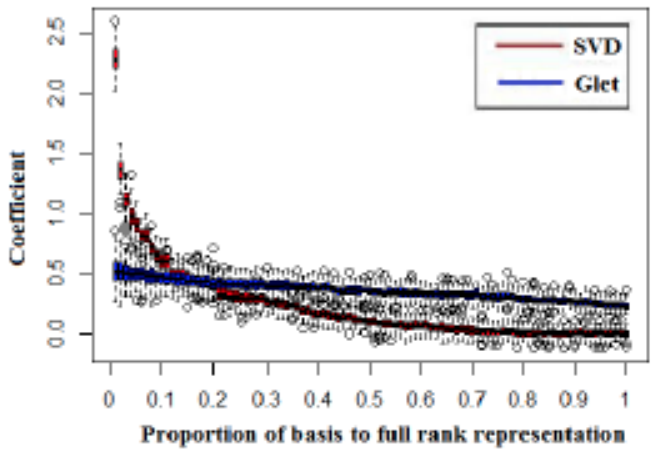
Analysis of criminal associations

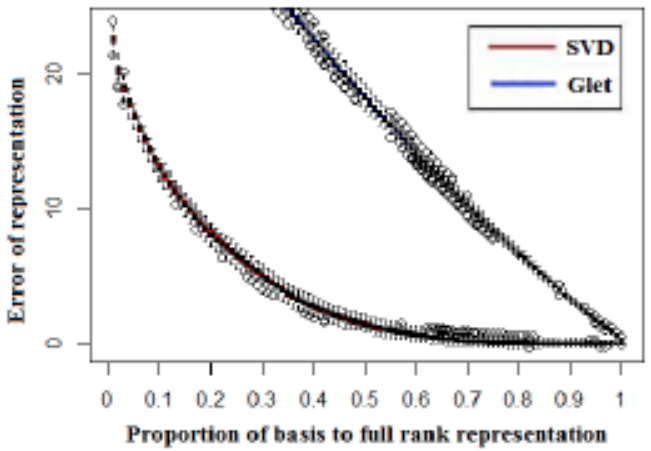


An implied notion of information









Remarks

- Properties

  Basis is interpretable – multiple/multi-scale membership

  Fast estimation algorithm – scales *O (sub no. edges)*

  Accurate – estimation algorithm learns structural zeros

  Theoretical guarantees

- Work in progress

  Imputing a fraction of structural zeros may be desirable

Agenda

- Graphlets
- Concluding remarks

Take home points

- Paired measurements raise new statistical problems where the familiar notions of sampling variability and sampling designs are challenged
- Graphlets

  Theory: complexity, redundancy, accuracy, scalability

  Practice: quantitative notion of social information

Acknowledgements and pointers

Facebook, AT&T. H Azari, P Balachandran, J Blitzstein, E Kolaczyk, C Marlow, XL Meng.

1. Graphlets: A semi-parametric method for analyzing large social and information with edge weights. *Artificial Intelligence and Statistics (AISTAT)*, 2012. Azari & Airoldi.

2. A survey of statistical network models. *Foundations & Trends in Machine Learning*, 2009. Goldenberg, Zheng, Fienberg & Airoldi.