Scan Statistics on Graphs

“Us” vs “Them”

Anomaly Detection in Streaming Graphs

Carey E. Priebe

Department of Applied Mathematics & Statistics
Johns Hopkins University, Baltimore, MD, USA

June 2012
Bristol, England

Collaborators

Heng Wang

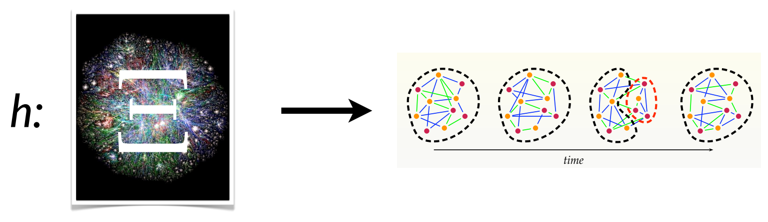
Youngser Park

Minh Tang

John Conroy, David Marchette, Andrey Ruhkin, Nam Lee, ...

Outline

Introduction



$$h_V : \Xi \to \mathcal{K}_V = \{red, orange\}$$

$$h_E : \Xi \to \mathcal{K}_E = \{green, blue\}$$

The map h provides a time series of (vertex- and edge-attributed) graphs

$$\{G_t\} = \{G(\Xi_t, h)\} = \{(V_t, E_t)\}$$

Motivation

Us:

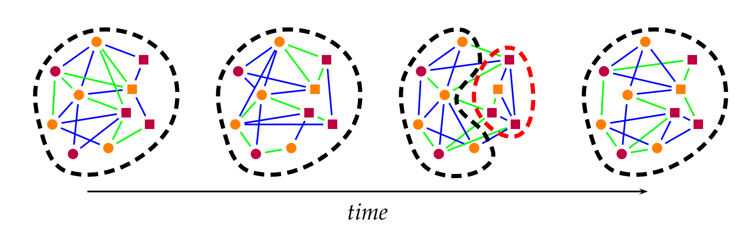
C.E. Priebe, J.M. Conroy, D.J. Marchette, Y. Park
"Scan Statistics on Enron Graphs", Computational & Mathematical Organization Theory, Vol 11, No 3, pp 229-247, 2005

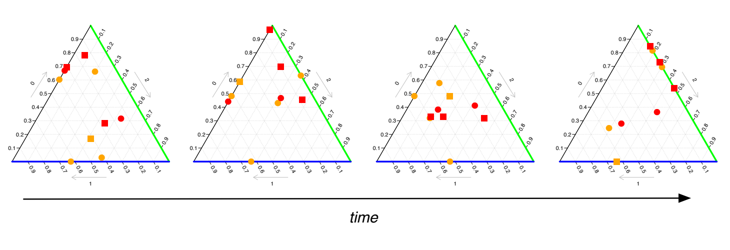
Them:

X. Wan, J. Janssen, N. Kalyaniwalla and E. Milios,
"Statistical analysis of dynamic graphs", Proceedings of AISB06: Adaptation in Artificial and Biological Systems, v3, pp.176-179, 2006.

X. Wan, N. Kalyaniwalla,
"Capturing causality in communications graphs", DIMACS/DyDAn Workshop on Computational Methods for Dynamic Interaction Networks, 2007.

A Latent Process Model
For Time Series of Attributed Graphs

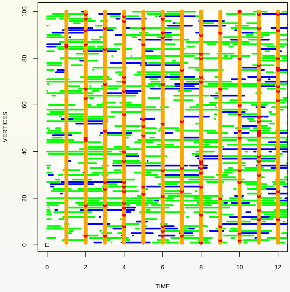


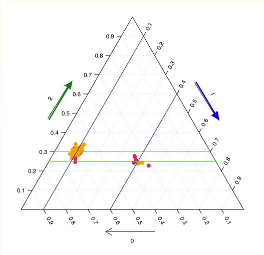


N.H. Lee and C.E. Priebe,
"A Latent Process Model for Time Series of Attributed Random Graphs", Statistical Inference for Stochastic Processes, Vol. 14, No. 3, pp. 231-253, 2011.

A Latent Process Model For Time Series of Attributed Graphs

Lee & Priebe (SISP, 2011)

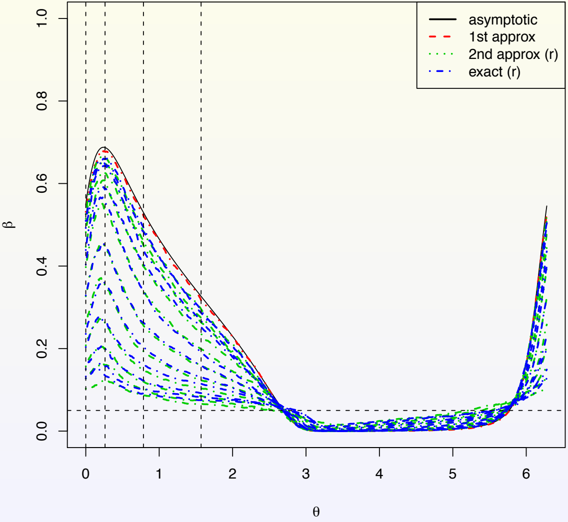




$$Q_i = \begin{pmatrix} -1 & & & 1 \\ & \ddots & & \vdots \\ & & -1 & 1 \\ \dfrac{\pi_{i,1}}{\pi_{i,d}} & \cdots & \dfrac{\pi_{i,d-1}}{\pi_{i,d}} & \dfrac{-\sum_{k=1}^{k=d-1} \pi_{i,k}}{\pi_{i,d}} \end{pmatrix}$$

A Latent Process Model For Time Series of Attributed Graphs

Lee & Priebe (SISP, 2011)



Scan Statistics

"moving window analysis" [1922 R.A. Fisher, 1965 J. Naus]:

to scan a small "window" (*scan region*) over data, calculating some *locality statistic* for each window; *e.g.*,

- number of events for a point pattern,
- average pixel value for an image.

scan statistic ≡ maximum of locality statistics:

If maximum of observed locality statistics is large, then the inference can be made that

there exists a subregion of excessive activity $\Longrightarrow$ *detection!*

Scan Statistics on Graphs

Let $G = (V, E)$ be a graph.

(k^{th}) neighborhood of v: $N_k[v; G] = \{w \in V : d(v, w) \leqslant k\}$, $k \geqslant 0$,

(k^{th}) scan region of v: $\Omega(N_k[v; G])$,

(k^{th}) locality statistic of v: $\Psi_k(v; G) = |E(\Omega(N_k[v; G]))|$,

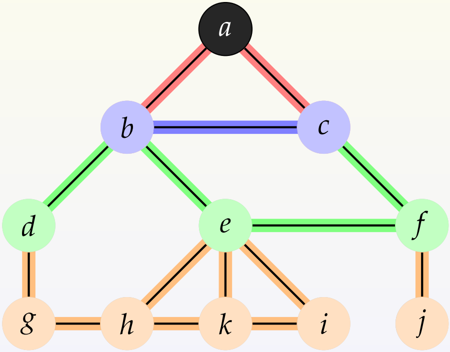
(k^{th}) scan statistic of G: $M_k(G) = \max_{v \in V(G)} \Psi_k(v; G)$.

"Maximum activity in k-neighborhood"

C.E. Priebe, J.M. Conroy, D.J. Marchette, Y. Park
"Scan Statistics on Enron Graphs", Computational & Mathematical Organization Theory, Vol 11, No 3, pp 229-247, 2005

Example of Scan Statistics

$|V| = 11, \ |E| = 15, \ \Psi_k(v; G) = |E(\Omega(N_k[v; G]))|$



k	$E(\Omega(N_k[\mathbf{a}; G]))$	$\Psi_k(\mathbf{a})$
0	$\ominus$	2
1	$\ominus + \ominus$	3
2	$\ominus + \ominus + \ominus$	7
3	$\ominus + \ominus + \ominus + \ominus$	15

	d	Ψ_1	Ψ_2	Ψ_3
a	2	3	7	15
b	4	5	14	15
c	3	4	8	15
d	2	2	8	14
e	5	7	15	15
f	3	3	12	15
g	2	2	7	14
h	3	4	8	15
i	2	3	7	15
j	1	1	3	12
k	3	5	10	15

Scan Statistics and Time Series

Let $\{G_t\}$, $t = 1, \ldots, t_{max}$, be a time series of graphs.

(k^{th}) scan region: $\Omega(N_k(v; G_t))$.

(k^{th}) locality statistic: $\Psi_k(v; G_t) = |E(\Omega(N_k(v; G_t)))|$.

(k^{th}) scan statistic: $M_k(G_t) = \max_{v \in V(G_t)} \Psi_k(v; G_t)$.

Us vs Them

Us:

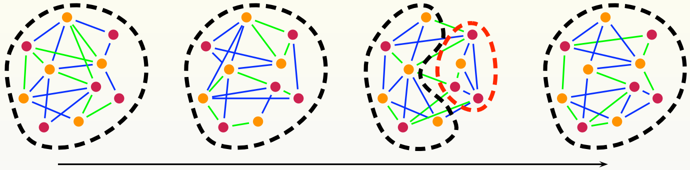
$$\Psi^{k}_{t;t'}(v) = |E(\Omega(N_k[v; G_{t'}]; G_{t'}))|$$

Them:

$$\Phi^{k}_{t;t'}(v) = |E(\Omega(N_k[v; G_{t}]; G_{t'}))|$$

$$t' \leqslant t$$

Normalization



vertex normalization:

$$\widetilde{J}^k_{t,\tau}(v) = \begin{cases} J^k_{t,t}(v) & \tau = 0 \\ \dfrac{J^k_{t,t}(v) - \hat{\mu}^k_{t,\tau}(v)}{\max(\hat{\sigma}^k_{t,\tau}(v), 1)} & \tau > 0 \end{cases}$$

$\hat{\mu}^k_{t,\tau}(v) = \frac{1}{\tau} \sum_{t'=t-\tau}^{t'=t-1} J^k_{t,t'}(v)$

$\hat{\sigma}^k_{t,\tau}(v) = \sqrt{\frac{1}{\tau - 1} \sum_{t'=t-\tau}^{t'=t-1} (J^k_{t,t'}(v) - \hat{\mu}^k_{t,\tau}(v))^2}$

Us vs Them

Us:

$$\Psi^k_{t;t'}(v) = |E(\Omega(N_k[v; G_{t'}]; G_{t'}))|$$

Them:

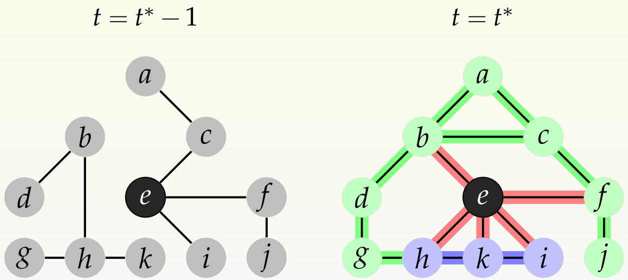
$$\Phi^k_{t;t'}(v) = |E(\Omega(N_k[v; G_t]; G_{t'}))|$$

$$t' \leqslant t$$

$$\hat{\mu}^k_{t,\tau}(v) = \frac{1}{\tau} \sum_{t'=t-\tau}^{t'=t-1} J^k_{t,t'}(v)$$

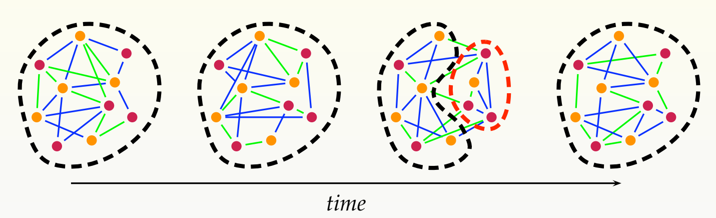
Normalization

$\tau = 1$



	$J^k_{t^*,t^*}(e)$		$\widehat{\mu}^k_{t^*,\tau}(e)$		$\widetilde{J}^k_{t^*,\tau}(e)$	
	$k=0$	$k=1$	$k=0$	$k=1$	$k=0$	$k=1$
us	5	7	3	3	2	4
them	5	7	2	4	3	3

Normalization



temporal normalization:

$$S^k_{t,\tau,\ell} = \begin{cases} \widetilde{M}^k_{t,\tau} = \max_v(\widetilde{J}^k_{t,\tau}(v)) & \ell = 0 \\ \dfrac{\widetilde{M}^k_{t,\tau} - \widetilde{\mu}^k_{t,\tau,\ell}}{\max(\widetilde{\sigma}^k_{t,\tau,\ell}, 1)} & \ell > 0 \end{cases}$$

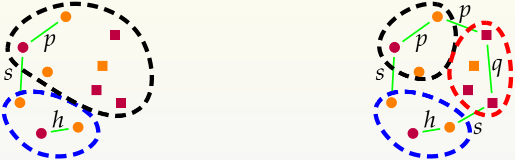
$\widetilde{\mu}^k_{t,\tau,l} = \frac{1}{l} \sum_{t'=t-l}^{t'=t-1} \widetilde{M}^k_{t',\tau}$

$\widetilde{\sigma}^k_{t,\tau,l} = \sqrt{\frac{1}{l-1} \sum_{t'=t-l}^{t'=t-1} (\widetilde{M}^k_{t',\tau} - \widetilde{\mu}^k_{t,\tau,l})^2}$

Heterogeneous Null

homogeneous null: $p < q$

heterogeneous null: $p \leqslant s < \{h, q\}$



Theory

$$H_0 : \{Q_v\} = \{Q_1^0(t), \cdots, Q_n^0(t)\},\ \forall t$$

$$H_A : \begin{cases} \{Q_v\} = \{Q_1^0(t), \cdots, Q_n^0(t)\},\ t < t^\star - 1 \\ \{Q_v\} = \{Q_1^A(t), \cdots, Q_m^A(t), Q_{m+1}^0(t), \cdots, Q_n^0(t)\},\ t \geqslant t^\star - 1 \end{cases}$$

1st approximation $\implies$ stochastic block model:

$$H_0 : G_t \overset{\text{iid}}{\sim} SBM(P_{B\times B}, n_{B\times 1}),\ \forall t$$

$$H_A : \begin{cases} G_t \overset{\text{iid}}{\sim} SBM(P_{B\times B}, n_{B\times 1}),\ t \leqslant t^\star - 1 \\ G_t \overset{\text{iid}}{\sim} SBM(P_{B\times B} + diag(0, \cdots, 0, \delta), n_{B\times 1}),\ t \geqslant t^\star \end{cases}$$

Theory

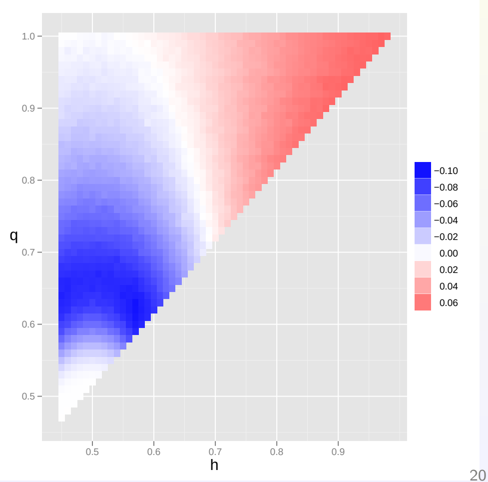
maxdegree, 1st approximation

Theorem

$$\lim_{n\to\infty} S \overset{\mathscr{L}}{=} \sum_{c=1}^{C_{T,H}} \pi_{T,H,c}\, g(\cdot;\theta_{T,H,c}).$$

g : *Gumbel*
C : *number of components*
π : *mixture coefficients*
$T \in \{\Psi, \Phi\}$
$H \in \{H_0, H_A\}$

$\beta(\text{us}) - \beta(\text{them})$
$(\ell = 0,\ \tau = 1)$



Theory

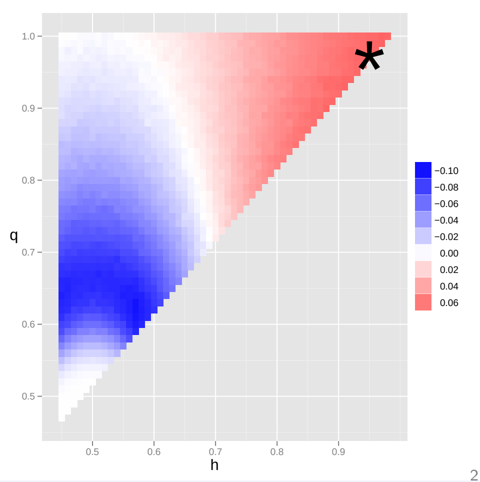
maxdegree, 1st approximation

Theorem

$$\lim_{n\to\infty} S \stackrel{\mathscr{L}}{=} \sum_{c=1}^{C_{T,H}} \pi_{T,H,c}\, g(\cdot;\theta_{T,H,c}).$$

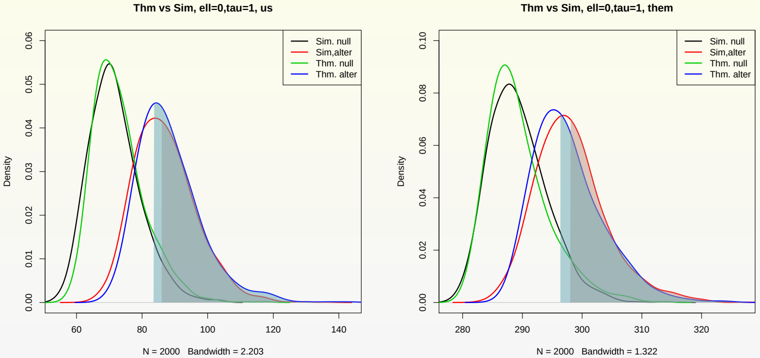
g : *Gumbel*
C : *number of components*
π : *mixture coefficients*
$T \in \{\Psi, \Phi\}$
$H \in \{H_0, H_A\}$

$\beta(\text{us}) - \beta(\text{them})$
$(\ell = 0,\ \tau = 1)$



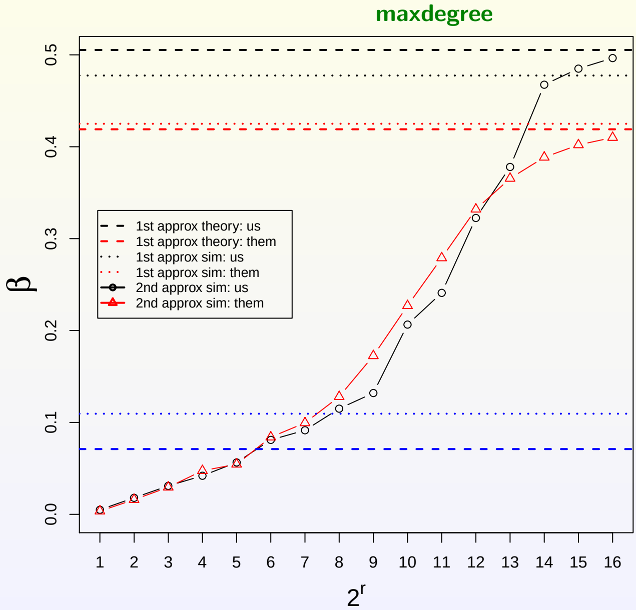
Theory & Simulation

maxdegree, 1st approximation



$\beta(\text{us})$		$\beta(\text{them})$
0.50	theory	0.42
0.48	simulation	0.43

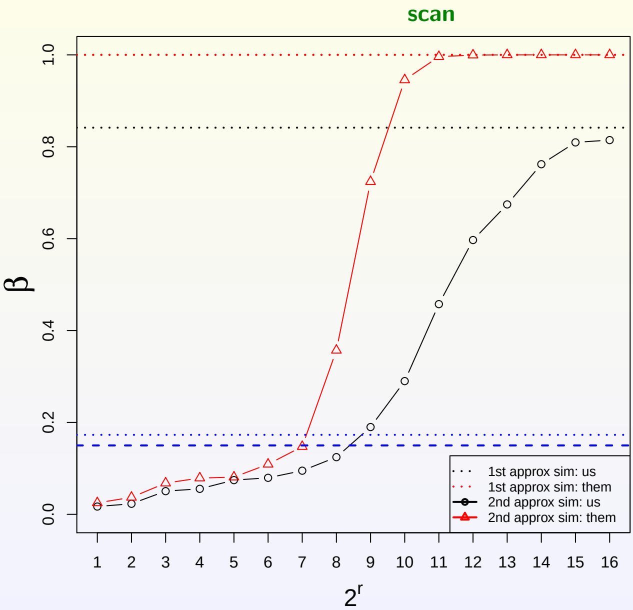
Theory & Simulation



$\tau = 1$

$\tau = 0$
(both
us & them)

Theory & Simulation



$\tau = 1$

$\tau = 0$
(both
us & them)

Conclusion

"them" is admissible!

Discussion

- there remains theory, simulation, and experiments yet to be done . . .
- *power vs. computational complexity tradeoff* for scan statistics on streaming graphs!

More References

http://www.cis.jhu.edu/~parky/CEP-Publications

- A. Rukhin and C.E. Priebe, "On the Limiting Distribution of a Graph Scan Statistic," *Communications in Statistics - Theory and Methods*, Vol. 41, No. 7, pp. 1151-1170, 2012.
- H. Pao, G.A. Coppersmith and C.E. Priebe, "Statistical Inference on Random Graphs: Comparative Power Analyses via Monte Carlo," *Journal of Computational and Graphical Statistics*, Vol. 20, No. 2, pp. 395-416, 2011.
- A. Rukhin and C.E. Priebe, "A Comparative Power Analysis of the Maximum Degree and Size Invariants for Random Graph Inference," *Journal of Statistical Planning and Inference*, Vol. 141, pp. 1041-1046, 2011.
- D.J. Marchette and C.E. Priebe, "Scan Statistics for Interstate Alliance Graphs," *Connections*, Volume 28, Issue 2, pp. 43-64, 2008.
- M. Tang, Y. Park, N.H. Lee, and C.E. Priebe, "Attribute fusion in a latent process model for time series of graphs," submitted, 2011.
- C.E. Priebe, G.A. Coppersmith, and A. Rukhin, "You say 'graph invariant,' I say 'test statistic' ", *ASA Sections on Statistical Computing Statistical Graphics, SCGN Newsletter*, 21, 2010.

Leopold Kronecker to Hermann von Helmholtz:

"The wealth of your practical experience
with sane and interesting problems
will give to mathematics
a new direction and a new impetus."

Kronecker

Helmholtz