

# Overview of UMass Research

W. Gong, D. Towsley

# Networks with MVHT distributions

- MVHT distributions ubiquitous in networks
  - in-degree, out-degree, reciprocated degree, labels, aggregate weights, ...

Q: How to model, generate, estimate, classify, learn network structures?  
How do networks evolve?

# Outline

- outline (D. Towsley)
- modeling, generating, estimating networks (Towsley)
- classifying networks, distributions (Gong)
- competition in growing networks (Jiang)
- learning networks from data (Atwood)

# Kaprivsky model [Krapivsky et al 2001]

□ with prob.  $p$

- new node attaches to existing node  $v$ , with prob. proportional to  $d_v^{\text{in}} + \lambda$

□ with prob.  $q = 1 - p$

- new edge  $u \rightarrow v$  connects existing nodes, with prob. proportional to  $(d_v^{\text{in}} + \lambda)(d_u^{\text{out}} + \mu)$

□ joint distribution

- when  $v_{\text{in}} = v_{\text{out}}$

$$\mathbb{P}(d_{\text{in}} = i, d_{\text{out}} = j) \sim C \frac{i^{\lambda-1} j^{\mu}}{(i+j)^{2\lambda+1}}$$

# Kaprivsky model

- efficient network generation algorithm - Atwood (UMass)
  - $O(n \log n)$
  - generates  $10^6$  node networks in seconds
  - node fitness, edge weights, other variants
- analysis - Gena
- parameter estimation - Jiang (UMass), Davis

# Kaprivsky model: limitations

- ❑ cannot control in/out degree correlation
- ❑ cannot directly account for reciprocity
- ❑ network datasets exhibit significant variations in both

More useful models?

# A versatile network model (UMass, UMN)

- generate undirected CA network
  - attach to  $i$  in proportion to  $\deg(i) + \lambda$ ,  $\lambda > -1$
- assign directions randomly
  - undirected, prob.  $p$
  - directed, prob.  $1 - p$ , each direction prob.  $(1 - p)/2$
- marginal unreciprocated in-, out-, reciprocated degree distr.

$$P(d_k = i) \propto i^{-3-\lambda}, \quad k = in, out, re$$

- asymptotic joint distribution

$$P(d_{in} = i, d_{out} = j, d_{re} = k) \sim \frac{C}{(i+j+k)^{3+\lambda}} \binom{i+j+k}{i \ j \ k} p^k \left(\frac{1-p}{2}\right)^{i+j}$$

# Ongoing work (UMass, UMN)

- explore other network models
  - model clustering (HT cluster sizes)
  
- place joint distributions in MRV framework
  - leverage Gena's recent work



# Complex Graph Similarity Testing and Multivariate Distribution Comparison Using Random Walks

Shan Lu, Jieqi Kang, Weibo Gong, Don Towsley

UMASS Amherst

# Outline of the approach

- Need for fast similarity testing among large data group/complex networks
- Existing algorithms are combinatorial in nature
- Small-time asymptotic results for diffusion on manifold motivated our approach
- Analogues on graphs/large data groups
- 1d experiments to seek understanding
- Analyzing graphs with 2d distributions
- Collaborations (Zhi-Li, Gena)
- Future work

Consider diffusion on Riemannian manifold  $M$ :

$$\Delta_M u = \frac{\partial u}{\partial t}, \quad t > 0,$$

with initial condition  $u(x, 0) = 1$

and boundary conditions  $u(x, t) = 0, \quad x \in \partial M, t > 0$

Using the  $\Delta_M$  spectral resolution  $\{\lambda_k, \phi_k\}$ ,  $\lambda_1 < \lambda_2 < \dots < \lambda_k < \dots$

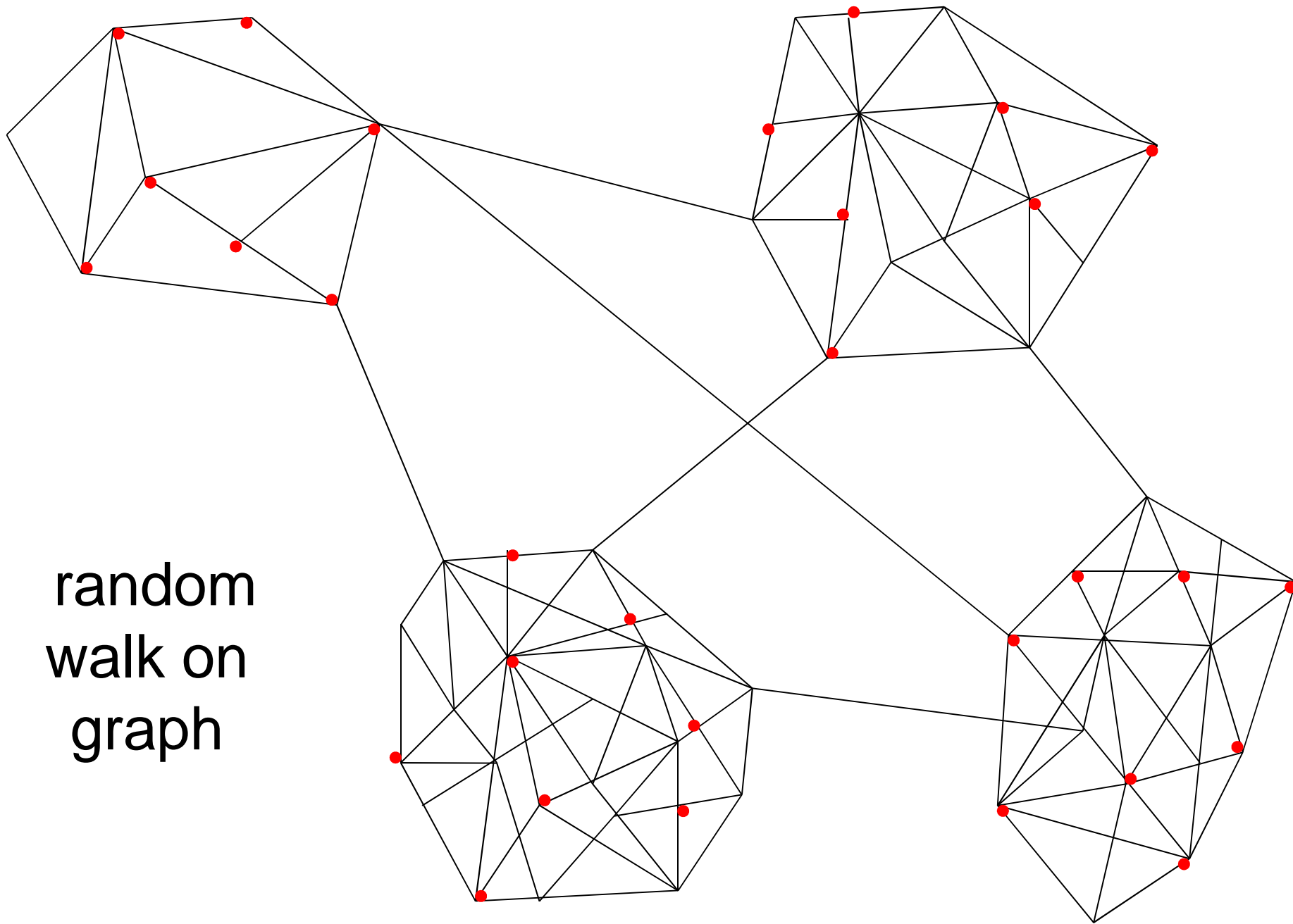
The Dirichlet heat kernel for  $M$  is  $p(x, y, t) = \sum_{k=1}^{\infty} e^{-\lambda_k t} \phi_k(x) \phi_k(y)$

and the heat content

$$h(t) = \int_M u(x, t) dx = \int_M \int_M p(x, y, t) dy dx = \sum_{k=1}^{\infty} e^{-t\lambda_k} \left( \int_M \phi_k(x) dx \right)^2$$

Note  $\int_M \phi_k(x) dx = \int_M \phi_k(x) u(x, 0) dx$  is the Fourier coefficient of the

initial condition in the eigen space spanned by  $\{\phi_1, \phi_2, \dots, \}$  !



random  
walk on  
graph

# Notations

*Laplacian matrix:*  $L = D - A$

*Normalized Laplacian:*  $\mathcal{L} = D^{-1/2} L D^{-1/2}$

*Random walk Laplacian:*  $L_r = D^{-1} L$

Let  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$  be the eigenvalues of  $\mathcal{L}$  and  $\phi_i, i = 1, \dots, N$  the corresponding eigenvectors. With  $\Lambda = \text{diag}[\lambda_i]$  and  $\Phi = [\phi_1, \dots, \phi_N]$ ,

$$\mathcal{L} = \Phi \Lambda \Phi^{-1} \quad L_r = (D^{-1/2} \Phi) \Lambda (\Phi^{-1} D^{1/2})$$

where  $\Phi^{-1} = [\pi_1; \dots; \pi_N]$ .

# Notations

*Partition vertex set  $V$  into two subsets, the set of all interior nodes  $iD$  and the set of all boundary nodes  $\partial D$ .*

*Label the interior vertices  $1, \dots, m$  and the boundary vertices  $m + 1, \dots, n$ . The normalized  $\mathcal{L}$  can be partitioned into:*

$$\mathcal{L} = \begin{bmatrix} \mathcal{L}_{iD,iD} & \mathcal{L}_{\partial D,iD} \\ \mathcal{L}_{iD,\partial D} & \mathcal{L}_{\partial D,\partial D} \end{bmatrix}$$

*Let the boundary vertices be absorbing in the heat equation in next page. Use  $\mathcal{L}$  for  $\mathcal{L}_{iD,iD}$  for convenience henceforth.*

# Heat Equation and Heat Content

**Heat Equation** is associated with the normalized graph Laplacian

$$\frac{\partial h_t}{\partial t} = -\mathcal{L}h_t \quad \Rightarrow \quad h_t = e^{-\mathcal{L}t}$$

with a given initial condition.

**Heat Content:**

$$Q(t) = \sum_{u \in iD} \sum_{v \in iD} h_t(u, v) = \sum_{u \in iD} \sum_{v \in iD} \sum_{i=1}^m e^{-\lambda_i t} \phi_i(u) \pi_i(v)$$

$$\alpha_i = \sum_{u \in iD} \sum_{v \in iD} \phi_i(u) \pi_i(v)$$

$$Q(t) = \sum_{i=1}^m \alpha_i e^{-\lambda_i t}$$

# Heat Equation and Heat Content

- Time derivatives of the heat content:

$$\dot{Q}(t) = - \sum_{i=1}^m \alpha_i \lambda_i e^{-\lambda_i t}$$

$$\ddot{Q}(t) = \sum_{i=1}^m \alpha_i \lambda_i^2 e^{-\lambda_i t}$$

- When  $t \rightarrow 0$ ,

$$Q(t)|_{t \rightarrow 0} = \sum_i \alpha_i = 1 \quad \dot{Q}(t)|_{t \rightarrow 0} = - \sum_i \alpha_i \lambda_i \quad \ddot{Q}(t)|_{t \rightarrow 0} = \sum_i \alpha_i \lambda_i^2$$



# Lazy Random Walk Approximation

1. *Transition matrix*  $M = D^{-1}A$ .
2. *Lazy random walk*  $M_L = (1 - \delta)I + \delta M$ .
3. For any given time  $t = k\delta$ , take the limit with  $k \rightarrow \infty$  ( $\delta \rightarrow 0$ ),

$$P_t = M_L^k P_0 = \left[ I - \frac{t}{k} L_r \right]^k P_0 \rightarrow e^{-L_r t} P_0$$

$$M_L^k(u, v) \rightarrow \sum_{i=1}^m e^{-\lambda_i t} \phi_i(u) \pi_i(v) \left( \frac{d_v}{d_u} \right)^{1/2}$$

$$\hat{Q}(t) = \sum_{u \in iD} \sum_{v \in iD} M_L^k(u, v) \left( \frac{d_u}{d_v} \right)^{1/2} \rightarrow Q(t)$$

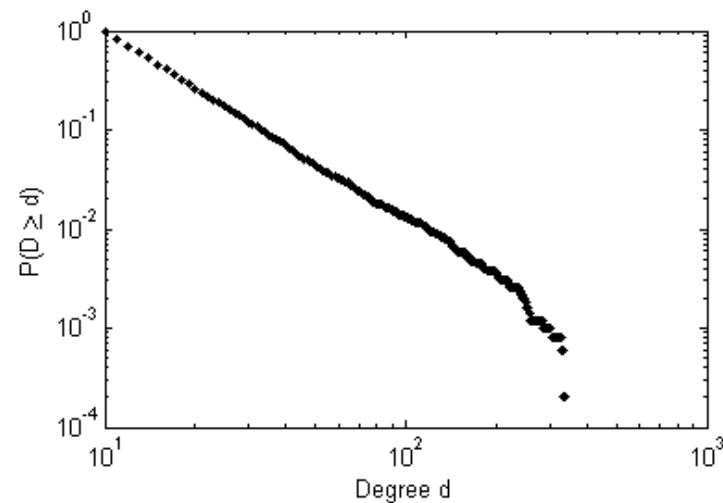
# Graph Similarity Testing

- Undirected Graphs
  - Barabási–Albert model vs. Erdős–Rényi model
- Directed Graphs
  - Krapivsky's model (2001)
  - Multivariate power law degree distribution
  - Krapivsky's model vs. Erdős–Rényi model

# Barabási–Albert vs. Erdős–Rényi Model

**Barabási–Albert model:** starts with  $s_0$  nodes; each new node is connected to  $s$  existing nodes with a probability proportional to the degree of the existing nodes.

Degree distribution follows  $P(D = d) \sim d^{-3}$ .

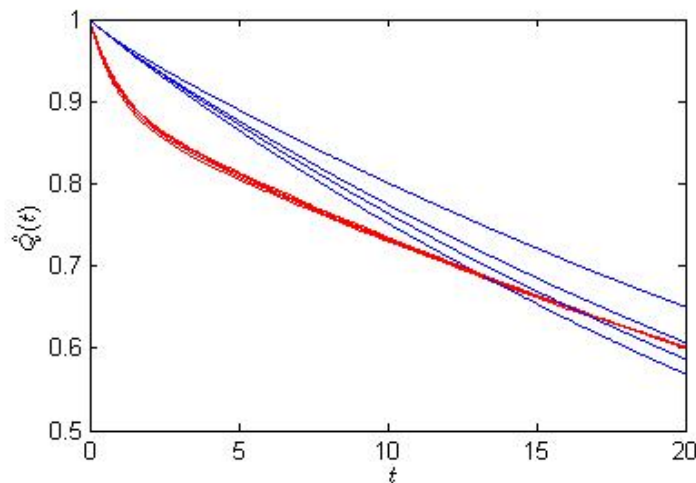
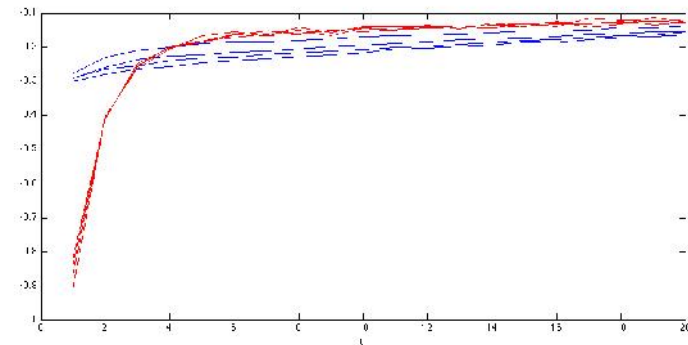


**Erdős–Rényi model:**  $G(n, p)$ , each edge is included in the graph with probability  $p$  independent from other edges.

# Barabási–Albert vs. Erdős–Rényi Model

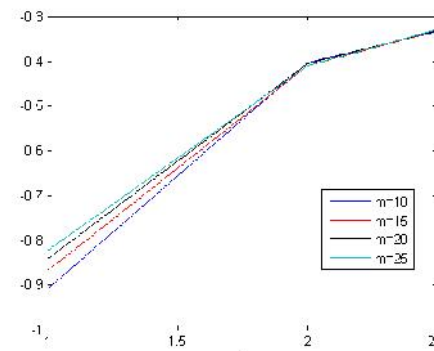
Two groups of graphs: power law graphs generated by B-A model and random graphs generated by E-R model. Average degree varies from 20 to 50. Heat content in each group are plotted in the same color.

Boundary selection: nodes with the smallest degrees.



- Power Law graphs generated by B-A model
- Random graphs generated by E-R model

The time derivatives of the heat contents for the two groups of graphs

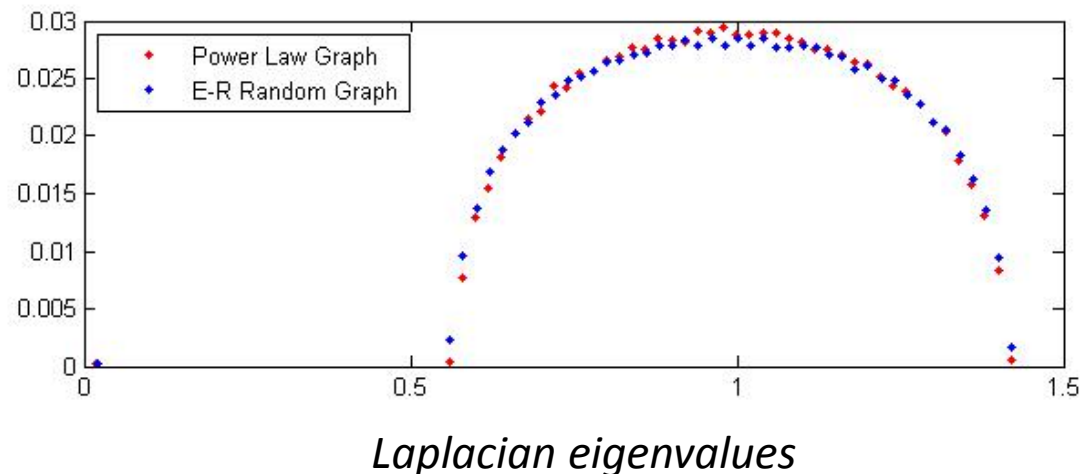


The initial time derivative of the heat contents for power law graphs with different mean degrees

# Barabási–Albert vs. Erdős–Rényi Model

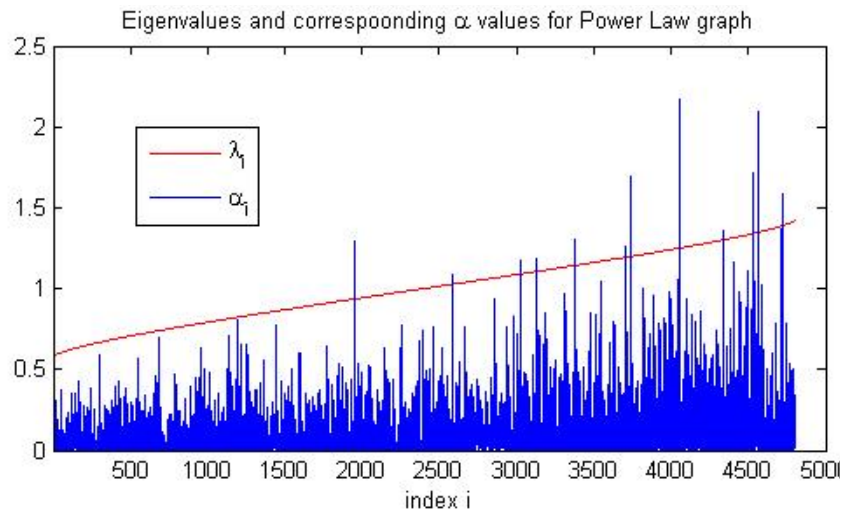
## *Laplacian Spectrum of two graphs with mean degree 20*

- The eigenvalues of the normalized Laplacian satisfy the semicircle law under the condition that the minimum expected degree is relatively large. (*Chung et. al. 2003*)



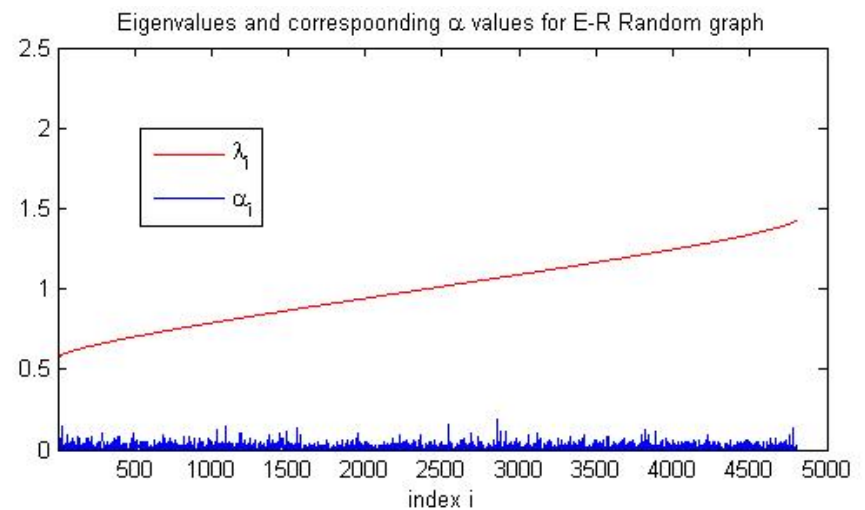
# Barabási–Albert vs. Erdős–Rényi Model

$$Q(t) = \sum_{i=1}^m \alpha_i e^{-\lambda_i t}$$



$$\lambda_1 = 0.0195 \quad \alpha_1 = 4260.2$$

Power Law graph generated by B-A model



$$\lambda_1 = 0.0210 \quad \alpha_1 = 4745.6$$

Random graph generated by E-R model

# Krapivsky vs. Erdős–Rényi Model

- Krapivsky's Model (*'Degree distributions of growing networks'*, 2001)
  - With probability  $p$ , a new node is introduced and attached to a target node  $u$  with probability proportional to  $d_u^{\text{in}} + \lambda_{\text{in}}$ .
  - With probability  $q = 1 - p$ , a new link from node  $v$  to node  $u$  is created with probability proportional to  $(d_u^{\text{in}} + \lambda_{\text{in}})(d_v^{\text{out}} + \lambda_{\text{out}})$ .

- Degree distribution

$$P(d^{\text{in}} = i) \sim i^{-v_{\text{in}}}$$

$$v_{\text{in}} = 2 + p\lambda_{\text{in}}$$

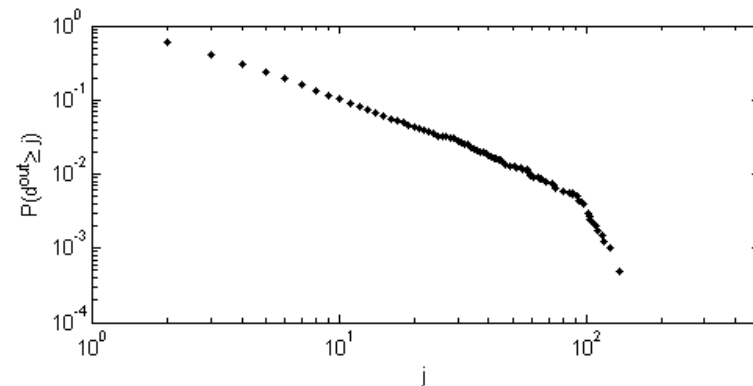
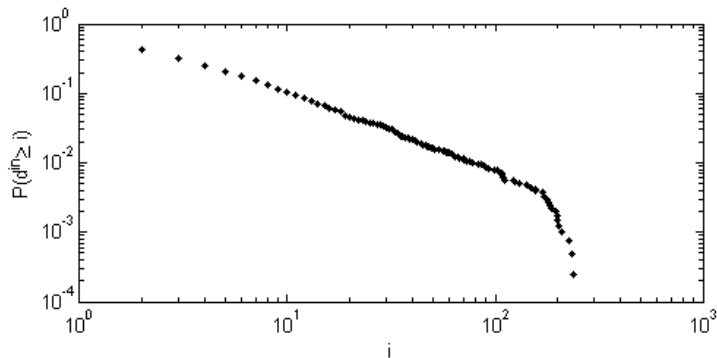
$$P(d^{\text{out}} = j) \sim j^{-v_{\text{out}}}$$

$$v_{\text{out}} = 1 + q^{-1} + p\lambda_{\text{out}}/q$$

- Average in-degree and out-degree:  $1/p$

# Krapivsky vs. Erdős–Rényi Model

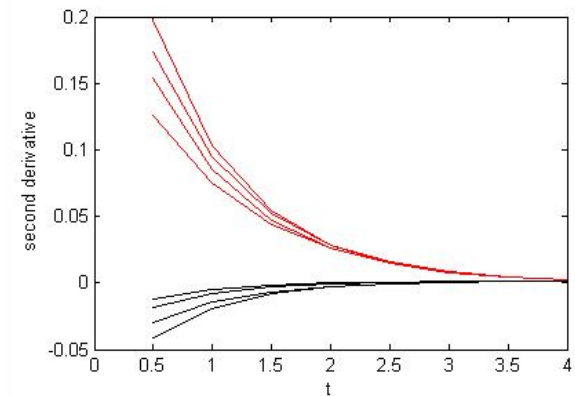
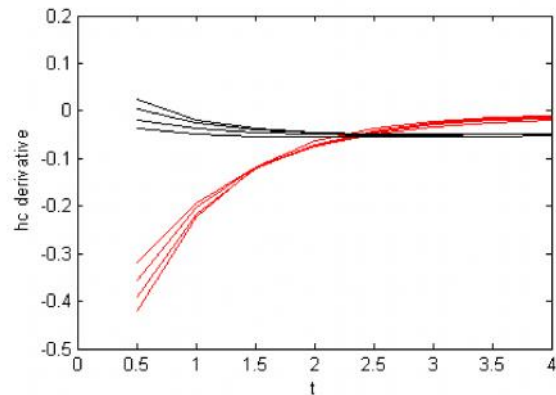
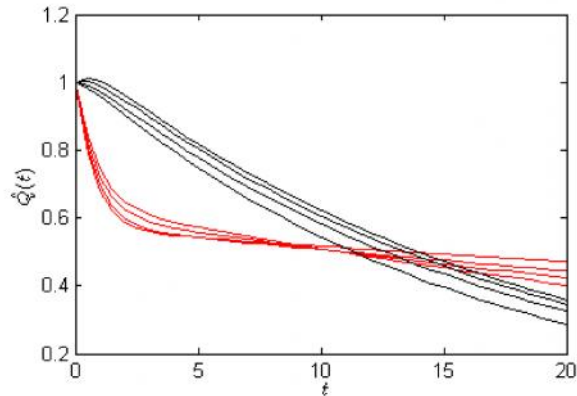
*Generated a 2000 nodes' directed graph using Krapivsky's model with  $p = 0.2$ ,  $\lambda_{in} = 2$  and  $\lambda_{out} = 1$ . CCDF of the in-degrees and out-degrees of the generated graph:*





# Krapivsky vs. Erdős–Rényi Model

Generated four 2000 nodes' directed graphs using Krapivsky's model with  $p=0.1, 0.15, 0.2$  and  $0.25$ , respectively. Also generated four directed graphs using E-R model with the same average degrees. Heat contents in each group are plotted in the same color.

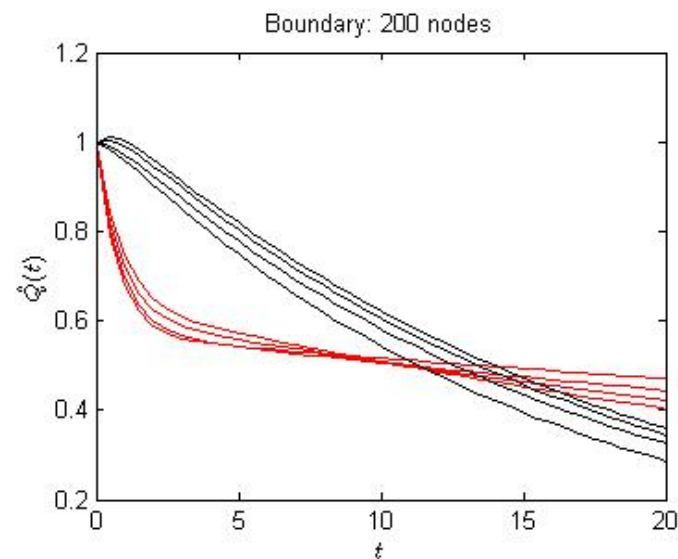
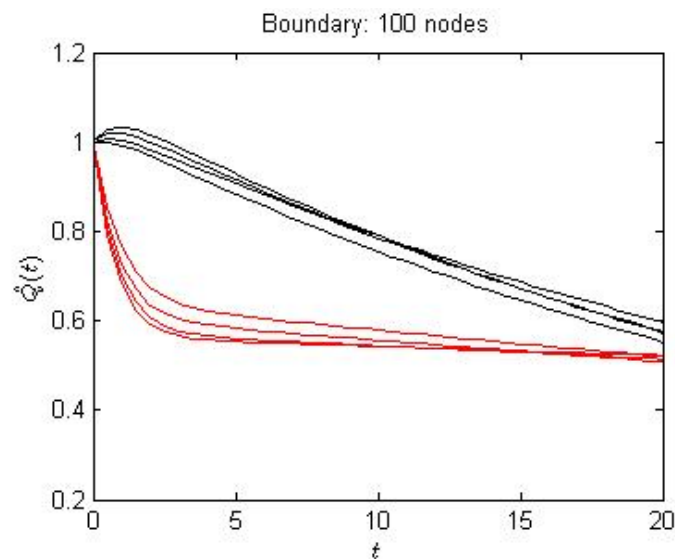


— Directed graphs with bivariate power Law degree distributions

— Directed E-R random graphs

# Krapivsky vs. Erdős–Rényi Model

*Boundary selection : nodes with smallest values of  $d^{\text{in}} \times d^{\text{out}}$*



*The number of boundary vertices impacts the heat contents of directed E-R random graphs more.*

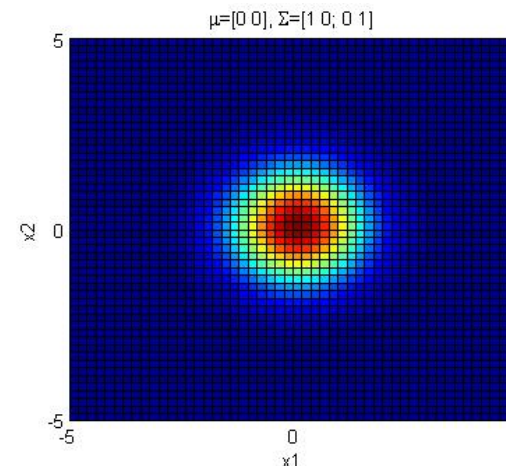
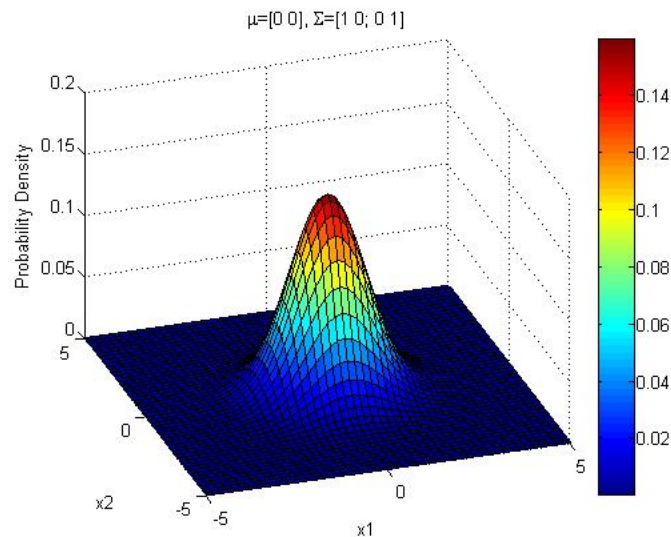
# Multivariate Distribution Comparison

- Multivariate Normal Distributions
  - Compare bivariate normal distributions with different covariance matrices
- Multivariate Power Law Distributions
  - Krapivsky's model, 2002
  - Correlated bivariate power law degree distributions
  - Correlated distributions vs. independent distributions

# Multivariate Normal Distribution

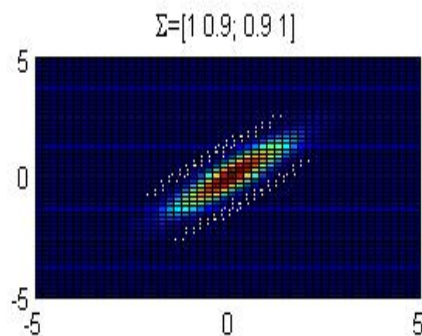
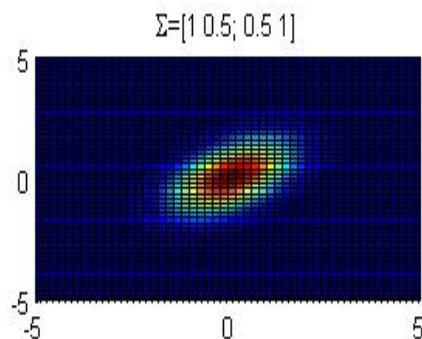
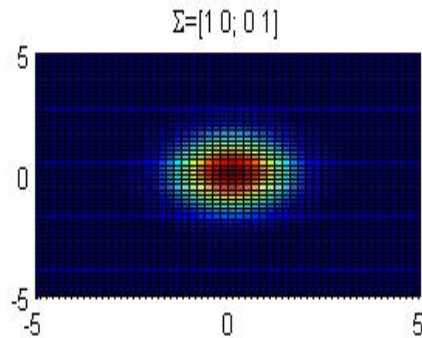
- Probability density function of the 2-dimensional bivariate normal distribution

$$y = f(x, \mu, \Sigma) = \frac{1}{\sqrt{|\Sigma|}(2\pi)^2} e^{-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)'}$$



$$W(p_1, p_2) = \frac{|f(p_1, \mu, \Sigma) - f(p_2, \mu, \Sigma)|}{\|p_1 - p_2\|^2}$$

# Multivariate Normal Distribution

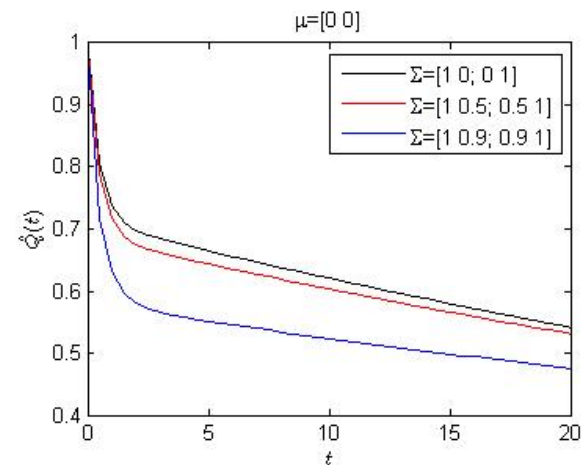


The Kullback–Leibler divergence from  $\mathcal{N}_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$  to  $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ , for non-singular matrices  $\boldsymbol{\Sigma}_0$  and  $\boldsymbol{\Sigma}_1$ , is:

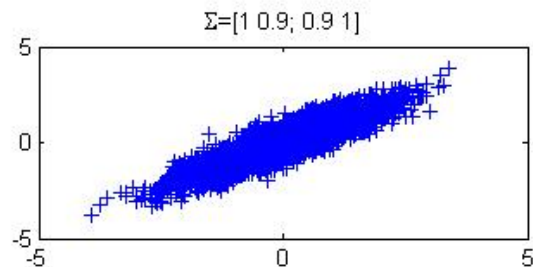
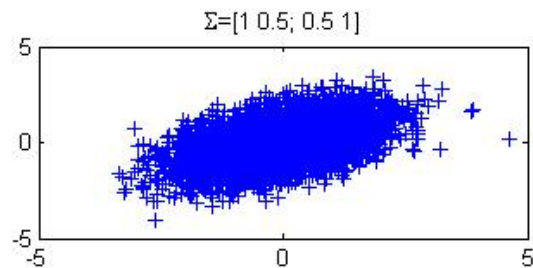
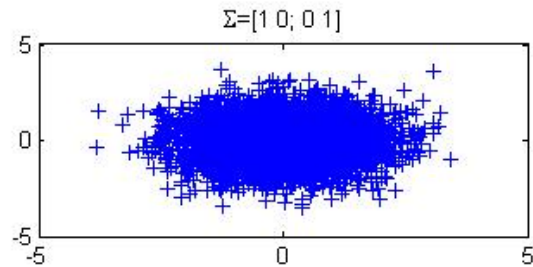
$$D_{\text{KL}}(\mathcal{N}_0\|\mathcal{N}_1) = \frac{1}{2} \left\{ \text{tr}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - K - \ln \frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} \right\},$$

where  $K$  is the dimension of the vector space.

$\mathcal{N}_0(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$ vs. $\mathcal{N}_1(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ ( $\boldsymbol{\mu} = [0\ 0]$ )	$D_{\text{KL}}(\mathcal{N}_1\ \mathcal{N}_0)$	$D_{\text{KL}}(\mathcal{N}_0\ \mathcal{N}_1)$	Symmetrised divergence
$\boldsymbol{\Sigma}_0 = [1, 0; 0, 1]$ vs. $\boldsymbol{\Sigma}_1 = [1, 0.5; 0.5, 1]$	0.1438	0.1895	0.3333
$\boldsymbol{\Sigma}_0 = [1, 0.5; 0.5, 1]$ vs. $\boldsymbol{\Sigma}_1 = [1, 0.9; 0.9, 1]$	0.4199	1.2082	1.6281
$\boldsymbol{\Sigma}_0 = [1, 0; 0, 1]$ vs. $\boldsymbol{\Sigma}_1 = [1, 0.9; 0.9, 1]$	0.8304	3.4328	4.2632



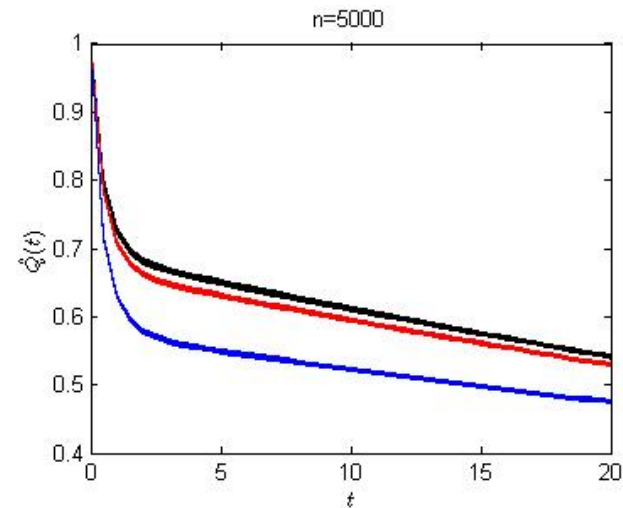
# Multivariate Normal Distribution



*20 samples for each distribution with 5000 random numbers in each sample.*

*Histogram is used for density estimation.*

*Heat contents for samples of the same distribution are plotted in the same color.*



# Multivariate distribution with power law marginal

- Krapivsky's Model (*'A statistical physics perspective on web growth'*, 2002) :
  - Some new node is introduced as isolated for webnet reality.

- Degree distribution

➤ *Marginal distribution*     $P(d^{\text{in}} = i) \sim i^{-v_{\text{in}}}$      $P(d^{\text{out}} = j) \sim j^{-v_{\text{out}}}$

$$v_{\text{in}} = 2 + p\lambda_{\text{in}}/q \qquad v_{\text{out}} = 2 + p\lambda_{\text{out}}/q$$

➤ *Joint Distribution (when  $\lambda_{\text{in}} = \lambda_{\text{out}} \equiv \lambda$ )*

$$P(d^{\text{in}} = i, d^{\text{out}} = j) \sim C \frac{(ij)^{\lambda-1}}{(i+j)^{1+\lambda+\lambda/q}}$$

*In-degree and out-degree of a node are correlated*

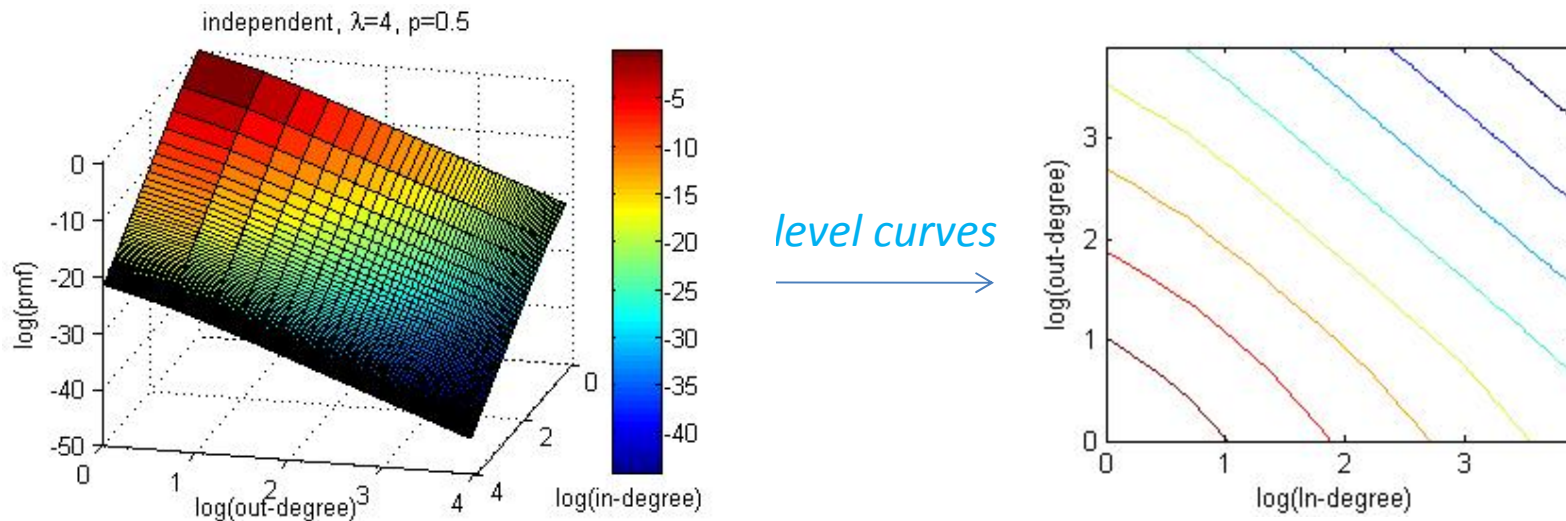
$$P(d^{\text{in}} = i, d^{\text{out}} = j) \neq P(d^{\text{in}} = i)P(d^{\text{out}} = j)$$

# Multivariate distribution with power law marginal

*Independent bivariate power law distribution with the same marginal distributions as in Krapivsky's model*

(  $\lambda_{in} = \lambda_{out} = 4$ ,  $p = 0.5$  in the figures below)

$$P(d^{in} = i, d^{out} = j) = P(d^{in} = i)P(d^{out} = j)$$

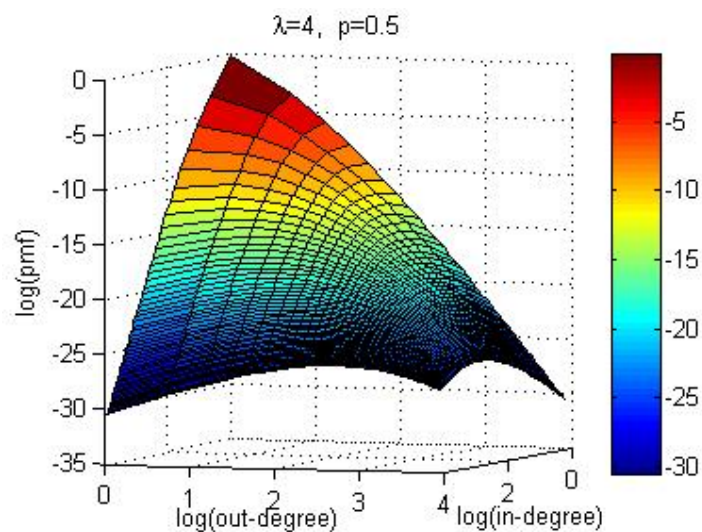




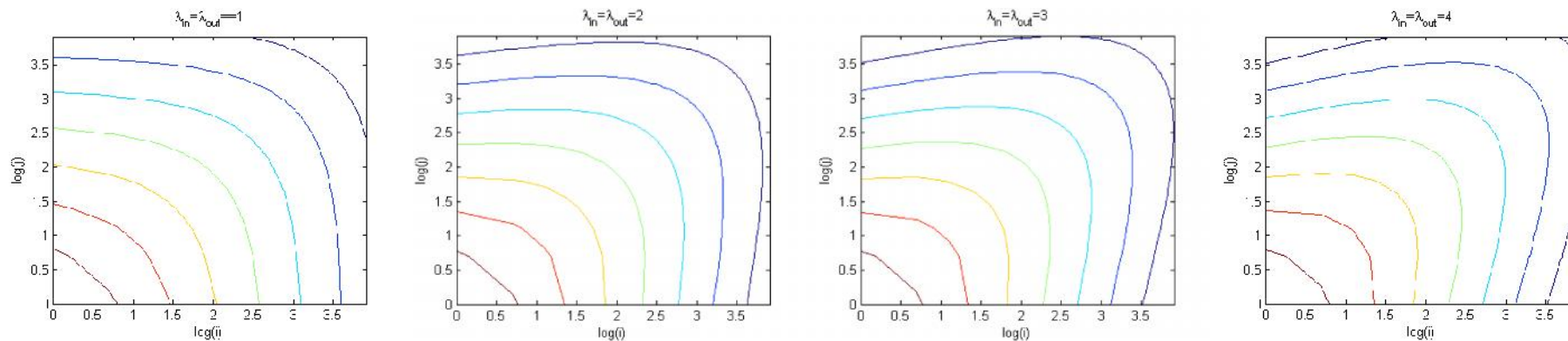
# Multivariate distribution with power law marginal

*The joint degree distribution generated by Krapivsky's model*

(  $\lambda_{in} = \lambda_{out} = 4$ ,  $p = 0.5$  in the figure on right)

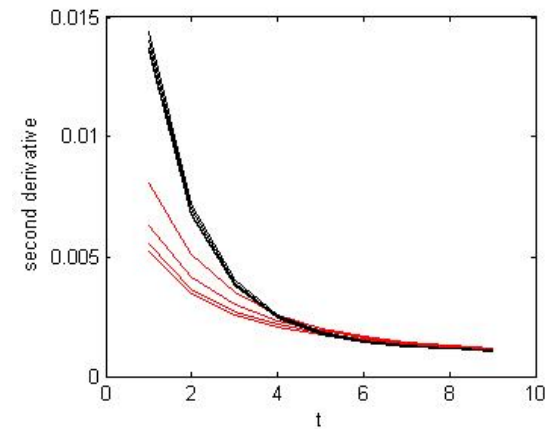
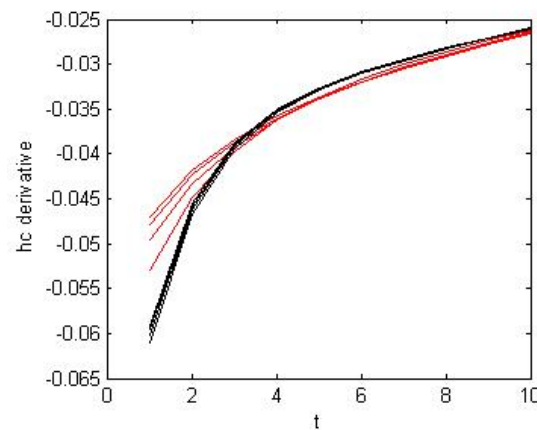
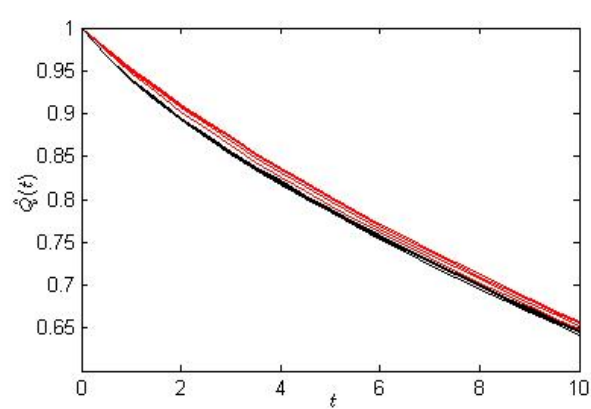


*The level curves for Krapivsky's distribution with different  $\lambda$  values*



# Multivariate distribution with power law marginal

*Heat contents and time derivatives of the two groups of distributions: distributions generated by Krapivsky's model with different  $\lambda$  values and the corresponding independent ones.*

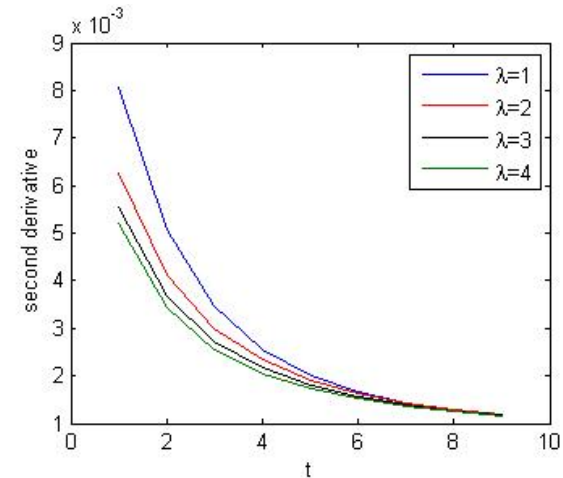
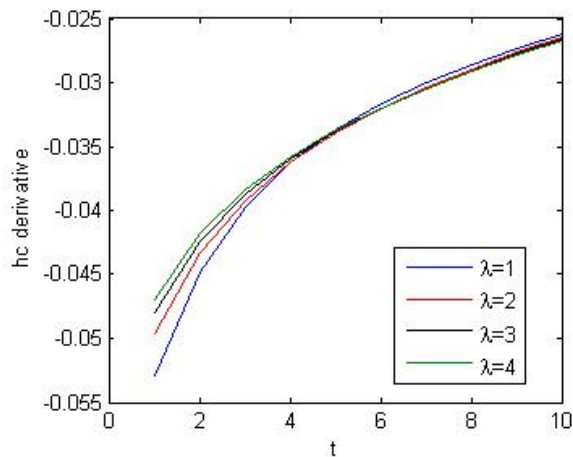


— Distributions generated by Krapivsky's model

— Independent bivariate power law distributions

# Multivariate distribution with power law marginal

*Heat content derivatives of the distributions generated by Krapivsky's model with different  $\lambda$  values.*



# Ongoing Work

- Mathematical definition and properties
  - L2 difference does not account for decreasing importance in time
- Graph Similarity Testing
  - *Consider other graph generative models*
  - *Consider real world network datasets*
- Multivariate Distribution Comparison
  - *Real data; higher dimension;*
  - *Theoretical understanding of correlation impact*
- Collaborations
  - UMASS-Cornell
  - UMASS-UMN