Complex Network Comparison Using Random Walks[*]

Shan Lu
Department of Electrical and Computer Engineering
University of Massachusetts Amherst
slu@ecs.umass.edu

Jieqi Kang
Department of Electrical and Computer Engineering
University of Massachusetts Amherst
jkang@ecs.umass.edu

Weibo Gong
Department of Electrical and Computer Engineering
University of Massachusetts Amherst
gong@ecs.umass.edu

Don Towsley
School of Computer Science
University of Massachusetts Amherst
towsley@cs.umass.edu

ABSTRACT

In this paper, we proposed a network comparison method based on the mathematical theory of diffusion over manifolds using random walks over graphs. We show that our method not only distinguishes between graphs with different degree distributions, but also different graphs with the same degree distributions. We compare the undirected power law graphs generated by Barabasi-Albert model and directed power law graphs generated by Krapivsky's model to the random graphs generated by Erdos-Renyi model. We also compare power law graphs generated by four different generative models with the same degree distribution.

Keywords

random walk; complex network; power law graph; graph comparison

1. INTRODUCTION

The asymptotic behavior of the heat content has been used as a tool to understand the geometry of a manifold domain [1, 15] or the connectivity structure of a graph [11, 12]. Heat content, as the solution of the heat equation associated with the Laplacian operator, summarizes the heat diffusion in the manifold domain or on the graph as a function of time. One property of the heat content method is that its asymptotic behavior as $t \to 0$ separates the heat content curves of different structures. This enables one to develop fast algorithms for comparing complex graphs. In [5, 6] it was pointed out that Monte-Carlo simulations of diffusions on graphs are effective in testing the similarity of complex graphs and that such simulations provide plausible mechanisms for many brain activities.

Graph comparison is a challenging task since graph sizes increase extremely fast in diverse areas. Many graph comparison methods have been proposed to quantitatively define the similarity between graphs. In [9] the authors summarize the existing methods into three categories: graph isomorphism, iterative methods, and feature extraction. The graph isomorphism and iterative methods are not scalable and thus not effective for large networks. Feature extraction methods extract features like degree distribution, eigenvalues to compare. These methods are closer in spirit to our method. However previously proposed features may not reflect the network connectivity structure very well. For example, in [12], the authors give an example where two isospectral non-isometric planar graphs can be distinguished by the heat content, despite the fact they share the same set of eigenvalues. In [7], the authors analysed the structural properties of graphs with the same degree distribution and found that different networks with the same degree distribution may have distinct structural properties. Using random walks to compare graphs is not a new idea. In [13], graphs are compared based on their mixing time, which measures the time needed for a random walk on the graph to approach the stationary distribution. This method may expect long walk length for the computation. Our method, on the other hand, focuses on the first few steps to compare and concerns the entire transient behavior of the random walk.

Our algorithm exhibits the following features. First, our method summarizes graph structure into a single time function so as to facilitate similarity testing. Second, the behavior of this function around time $t = 0$ forms the basis for the comparison, so that we can greatly reduce the computation time. Third, we use a lazy random walk to estimate the heat content function, thereby avoid computing the eigenvalues and eigenvectors of the graph Laplacian while retaining the spectral information. Finally we note that our method is robust to minor changes in large graphs according to the interlacing theorem in [2]. With these features, our algorithm is capable of handling very large complex networks.

The rest of the paper is organized as follows. In Section 2, we give the notations and review the concept of heat equa-

[*]This work is supported in part by the United States National Science Foundation CNS-1065133 and CNS-1239102, and Army Research Office Contract W911NF-08-1-0233 and W911NF-12-1-03.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579363.

tion and heat content for graphs. In Section 3, we use the lazy random walk simulation method to estimate the heat content. In Section 4, the graph generative models used in experiment part are introduced. Experiment settings and results are presented in Section 5. Section 6 summarizes the main results and discusses future work.

2. HEAT EQUATION AND HEAT CONTENT

2.1 Notations

Let $G = (V, E)$ denote a graph with vertex set V and edge set $E \subseteq V \times V$ with adjacency matrix $A = [a_{uv}]$. $a_{uv} = 1$ if there is an edge from u to v; otherwise $a_{uv} = 0$. The out-degree matrix $D = diag[d_u]$ with $d_u = \sum_v a_{uv}$. The graph Laplacian of a graph is defined as $L = D - A$ and the normalized Laplacian is defined as $\mathcal{L} = D^{-1/2}LD^{-1/2}$ [3]. With the random walk Laplacian $L_r = D^{-1}L$, we have $L_r = D^{-1/2}\mathcal{L}D^{1/2}$.

Without loss of generality, we assume that the Laplacian L is diagonalizable and hence $\mathcal{L}$ is diagonalizable. Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of $\mathcal{L}$ and $\phi_i, i = 1, \cdots, n$ the corresponding eigenvectors. With $\Lambda = diag[\lambda_i]$ and $\Phi = [\phi_1, \cdots, \phi_n]$ we can diagonalize $\mathcal{L}$ as $\mathcal{L} = \Phi\Lambda\Phi^{-1}$, where $\Phi^{-1} = [\pi_1; \pi_2; \cdots; \pi_n]$. Meanwhile

$$L_r = (D^{-1/2}\Phi)\Lambda(D^{-1/2}\Phi)^{-1}. \tag{1}$$

L_r and $\mathcal{L}$ share the same set of eigenvalues. $\mathcal{L}$ is the normalized graph Laplacian used in the heat equation on a graph. We use the relationship between $\mathcal{L}$ and L_r to develop a random walk simulation method in the later section.

2.2 Heat equation and heat content

Vertex set V is partitioned into two subsets, the set of all interior nodes iD and the set of all boundary nodes ∂D; $V = iD \cup \partial D$. The heat equation associated with the normalized graph Laplacian is

$$\begin{cases} \frac{\partial H_t}{\partial t} = -\mathcal{L}H_t \\ H_t(u, v) = 0 \text{ for } u \in \partial D, \end{cases} \tag{2}$$

with initial condition $H_0(u, u) = 1$ if $u \in iD$.

Let N denote the total number of vertices and n the number of interior vertices; then H_t is an $N \times N$ matrix. $H_t(u, v)$ measures the amount of heat that initiates from vertex u and ends up at vertex v at time t. All heat that flows to the boundary vertices is absorbed. We label the interior vertices $1, \cdots, n$ and the boundary vertices $n + 1, \cdots, N$. The normalized Laplacian $\mathcal{L}$ can be partitioned into four parts. The part related to the interior domain is denoted as $\mathcal{L}_{iD,iD}$. Since we are only interested in the heat remaining in the interior domain, we define the $n \times n$ matrix h_t with $h_t(u, v) = H_t(u, v)$ (for $u, v \in iD$). The solution to the heat equation is $h_t = e^{-\mathcal{L}_{iD,iD}t}$. For convenience, we slightly abuse notation and use Λ and Φ as the eigenvalue matrix and eigenvector matrix of $\mathcal{L}_{iD,iD}$:

$$h_t(u, v) = \sum_{i=1}^{n} e^{-\lambda_i t}\phi_i(u)\pi_i(v). \tag{3}$$

The heat content $Q(t)$ is defined as:

$$Q(t) = \sum_u \sum_v \sum_{i=1}^{n} e^{-\lambda_i t}\phi_i(u)\pi_i(v). \tag{4}$$

Letting $\alpha_i = \sum_u \sum_v \phi_i(u)\pi_i(v)$ yields

$$Q(t) = \sum_{i=1}^{n} \alpha_i e^{-\lambda_i t}. \tag{5}$$

We can also use the derivatives of the heat content $\dot{Q}(t) = -\sum_{i=1}^{m} \alpha_i\lambda_i e^{-\lambda_i t}$ for comparison to emphasize larger eigenvalues.

3. RANDOM WALK METHODS FOR HEAT CONTENT ESTIMATION

Computing eigenvalues and eigenvectors of the Laplacian matrix needed for evaluating the heat content is very time consuming for large complex networks. We consider a random walk where the walker moves from vertex u to a neighboring vertex v with probability a_{uv}/d_u. Define the transition matrix $M = D^{-1}A$ and the lazy random walk transition matrix as $M_L = (1 - \delta)I + \delta M$. For any given time $t = k\delta$, we have

$$P_t = M_L^k P_0 = [I - \frac{t}{k}L_r]^k P_0 \ \rightarrow \ e^{-L_r t}P_0. \tag{6}$$

Here the arrow ($\rightarrow$) implies taking the limit as $k \rightarrow \infty$ (at the same time $\delta \rightarrow 0$ while keeping $k\delta = t$). P_0 is the initial distribution of a random walker. We have $M_L^k \rightarrow e^{-L_r t}$. From equations (1), (3), and (4), we obtain the approximation for $Q(t)$:

$$\hat{Q}(t) = \sum_{u \in iD} \sum_{v \in iD} M_L^k(u, v)\sqrt{\frac{d_u}{d_v}}. \tag{7}$$

With the lazy random walk approximation, our algorithm avoids the computation of the eigenvalues and the eigenvectors. Instead of computing $M_L^k(u, v)$ with matrix multiplications, we can use the Monte Carlo method to estimate $M_L^k(u, v)$. The variance of the estimated value is inversely proportional to the amount of random walkers. Therefore, our method provides a trade off between precision and computation time.

4. GENERATIVE MODELS

We consider the following generative models for complex graphs.

Erdos-Renyi (E-R) model

The graph is constructed by connecting nodes randomly and independently. An edge is added to each pair of vertices with a given probability. Directed E-R graphs can also be generated using a similar mechanism.

Barabasi-Albert (B-A) model

Start with m_0 initial nodes. Each new node is connected to $m (m \leq m_0)$ existing nodes with a probability proportional to the number of links that the existing nodes already have. The degree distribution follows $P(D = d) \sim d^{-3}$.

Krapivsky's model for directed graphs

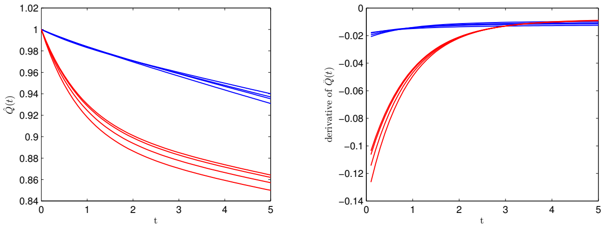
In [10], a graph generative model is proposed to describe growing processes in the Web Graphs (WG). With probability p, a new node is introduced and immediately attaches to an existing node u with probability proportional to $d_u^{\text{in}} + \lambda_{\text{in}}$, where d_u^{in} is the in-degree of node u. With probability q, a new edge from existing node v to node u is created with probability proportional to $(d_u^{\text{in}}+\lambda_{\text{in}})(d_v^{\text{out}}+\lambda_{\text{out}})$. This model produces directed graphs with marginal in-degree and out-degree distributions that are both heavy tailed. Let $P(d^{\text{in}} = i) \sim i^{-v_{\text{in}}}$ and $P(d^{\text{out}} = j) \sim j^{-v_{\text{out}}}$. We have $v_{\text{in}} = 2 + p\lambda_{\text{in}}$ and $v_{\text{out}} = 1 + q^{-1} + p\lambda_{\text{out}}/q$.

In [7], the authors used the following four models to generate graphs with the same degree distribution. The first assigns a degree to each vertex. The Molloy-Reed Model (M-R Model)[14] randomly connects a pair of vertices with probability proportional to the number of open connections. The Kalisky Model [8] starts from the vertex with the maximal degree and exhausts its open connections by randomly connecting it to other vertices. These vertices are the first layer vertices. The second layer vertices are selected by randomly connecting the remaining open connections in the first layer. Repeat this until there is no open connection. Model A and model B are new methods proposed in [7]. Model A randomly connects the open connections of the maximal degree node in each step to the available vertices until there is no open connections. Model B is the same as model A, except that the vertices connecting to the maximal degree node are selected in sequence according to a given vertices list.

5. EXPERIMENTAL RESULTS

5.1 Graphs with different degree distributions

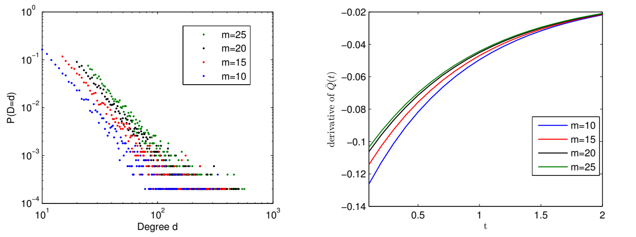
Two groups of graphs are generated using the B-A and E-R models respectively. The total number of nodes is 2000. Each group includes four graphs with average degree varying from 20 to 50. Boundary vertices are defined to be the 40 vertices with the smallest degrees. As shown in Fig. 1(a), the heat contents of the two groups of graphs follow different patterns. When t is close to zero, the heat contents for power law graphs drops faster than for E-R random graphs, but the decrease speed slows down once $t > 5$. The difference between the heat contents of these two types of graphs is illustrated more clearly if we focus on the time derivative of the heat content (Fig. 1(b)). When we compare the heat content derivatives for the power law graphs, the derivatives at the beginning part are in the order of the average degrees (as shown in Figure 2(b)).



(a) Heat Content (b) Heat Content Derivatives

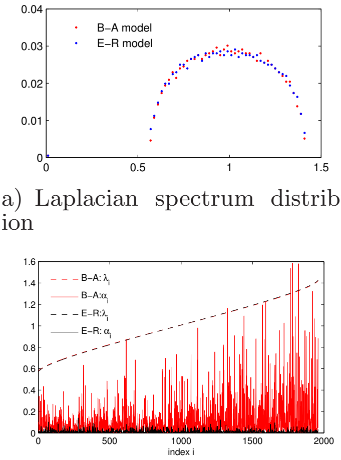
Figure 1: Undirected graph comparison (red: power law graphs; blue: E-R random graphs)

For the spectra of these two kinds of graphs, Chung *et.al.*[4] proved that eigenvalues of the normalized Laplacian for both E-R random graphs and power law graphs satisfy the semicircle law. The circle radius is almost the same for graphs with the same mean degree (as shown in Fig. 3(a)). Using only the Laplacian spectrum we can hardly distinguish the two types of graphs. However, according to equation (5), the values of α_i also play an important role in the heat contents. As shown in Fig. 3(b), the strengths (α) for the power law graph are much larger than those for the E-R random graph, which explains the different heat content behaviors for the two kinds of graphs.



(a) Degree Distributions (b) Heat Content Derivatives

Figure 2: Degree distributions and heat content derivatives for power law graphs



(a) Laplacian spectrum distribution

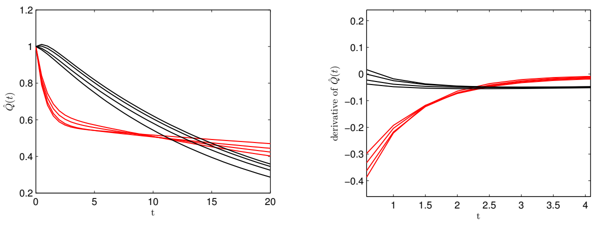
(b) Eigenvalues and corresponding α values

Figure 3: The Laplacian spectrum of one power law graph and one random graph with mean degree 20

For directed graphs comparison, two groups of graphs are generated using 'WG' model and the E-R model respectively. Each group contains four graphs with different average degrees by setting p to be 0.1, 0.15, 0.2 and 0.25, respectively. Boundary vertices are defined to be the 200 vertices with the smallest in-degree out-degree products (vertices with zero in-degrees are not candidates for boundaries). As shown in Fig. 4, the directed power law graphs and E-R random graphs exhibit similar behavior to undirected graphs.

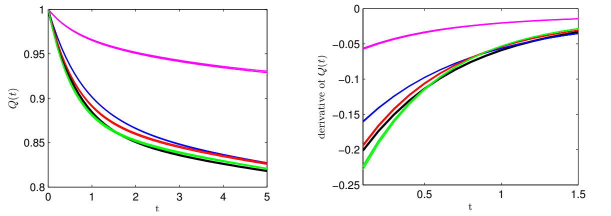
5.2 Graphs with the same degree distribution

We first generate a 2000 nodes power law graph using B-A model with $m = 2$. Then using each one of the 4 generative models, we independently generate 3 graphs with the same degree distribution. The heat contents and derivatives of all 13 graphs are shown in Figures 5(a) and 5(b).



(a) Heat content (b) Heat content derivative

Figure 4: Directed graph comparison (red line: power law graphs; black line: E-R random graphs)



(a) Heat content (b) Heat content derivative

Figure 5: Comparing graphs with the same degree distribution (Black: B-A; Blue: M-R; Red:Kalisky; Green: model A; Magenta: model B)

We observe from the results that graphs with the same degree distribution can be distinguished according to their heat content behaviors. The heat contents for graphs generated by the same model are clustered. And at the same time, although with the same degree distributions, the differences of the heat contents between the 5 generative models are also noticeable. We also notice that the heat contents for model B (the curves in color magenta) perform differently from the other four models (Molloy-Reed model, Kalisky model, model A and the B-A model). This result is consistent with the conclusion in [7] that, although with the same degree distribution, model B gives decentralized network with a larger number of components and a smaller giant component comparing to the other four models.

6. CONCLUSION

In this paper, we proposed a random walk method to estimate the heat content on graphs for the purpose of determining if two graphs are similar or not. We first apply the method to compare graphs with different degree distributions. Graphs with heavy tailed degree distribution have different heat content curves comparing to the random graphs generated by the E-R model: the decrease rate for the previous is much larger than that for the later at the very beginning part. Our method can also distinguish graphs with the same degree distribution but different structural properties. Experiments show that, our algorithm is better in graph comparison than some other feature extraction methods like eigenvalues and degree distributions. In our future work, we will apply our method to more general problems in graph comparison. We will also consider real world network datasets.

7. REFERENCES

[1] M. Vandenberg and P. B. Gilkey. Heat content asymptotics of a riemannian manifold with boundary. *Journal of Functional Analysis*, 120:48–71, 1994.

[2] S. Butler. Interlacing for weighted graphs using the normalized laplacian. *Electronic Journal of Linear Algebra*, 16:90–98, 2007.

[3] F. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[4] F. Chung, L. Lu, and V. Vu. Spectra of random graphs with given expected degrees. *Proceedings of the National Academy of Sciences of the United States of Amherica*, 100(11):6313–6318, 2003.

[5] W. Gong. Can one hear the shape of a concept? In *Proceedings of the 31st Chinese Control Conference (Plenary Lecture)*, pages 22–26, Hefei, China, July 2012.

[6] W. Gong. Transient response functions for graph structure addressable memory. In *Proceedings of the 52th IEEE Conference on Decision and Control*, Florence, Italy, December 2013.

[7] J.H.H.Grisi-Filho, R. Ossada, F. Ferreira, and M. Amaku. Scale-free networks with the same degree distribution: Different structural properties. *Physics Research International*, 2013:234180(1–9), 2013.

[8] T. Kalisky, R. Cohen, D. Ben-Avraham, and S. Havlin. Tomography and stability of complex networks. *Complex Networks*, 650:3–34, 2004.

[9] D. Koutra, A. Parikh, A. Ramdas, and J. Xiang, "Algorithms for graph similarity and subgraph matching," 2011, Available at `https://www.cs.cmu.edu/~jingx/docs/DBreport.pdf`.

[10] P. L. Krapivsky, G. J. Rodgers, and S. Redner. Degree distributions of growing networks. *Physical Review Letters*, 86:5401–5404, 2001.

[11] P. Mcdonald and R. Meyers. Diffusion on graphs, poisson problems and spectral geometry. *Transaction of the American Mathematical Society*, 354(12):5111–5136, 2002.

[12] P. Mcdonald and R. Meyers. Isospectral polygons, planar graphs and heat content. *Proceedings of the American Mathematical Society*, 131(11):3589–3599, 2003.

[13] A. Mohaisen, A. Yun, and Y. Kim. Measuring the mixing time of social graphs. *Internet Measurement Conference*, pages 383–389, 2010.

[14] M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing*, 7(03):295–305, 1998.

[15] J. Park and K. Kim. The heat energy content of a riemannian manifold. *Trends in Mathematics, Information Center for Mathematical Sciences*, 5(2):125–129, 2002.