Linear Algebra and its Applications 438 (2013) 2331–2338

Contents lists available at SciVerse ScienceDirect

Linear Algebra and its Applications

journal homepage: www.elsevier.com/locate/laa

A new dissimilarity measure for comparing labeled graphs

Nicolas Wicker [a,*], Canh Hao Nguyen [b], Hiroshi Mamitsuka [b]

[a] Laboratoire Painlevé, Université de Lille 1, 59655 Villeneuve d'Ascq, France
[b] Bioinformatics Center, ICR, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

ARTICLE INFO

Article history:
Received 6 June 2012
Accepted 6 October 2012
Available online 30 November 2012

Submitted by Shaun Fallat

AMS classification:
05C50
15A18

Keywords:
Labeled graph dissimilarity
Spectral graph comparison
Graph Laplacian

ABSTRACT

We use spectral graph theory to compare graphs that share the same node set, taking into account global graph structures. We propose a general framework using eigendecomposition of graph Laplacians. We show its special cases and propose a new dissimilarity measure that avoid problems of spectral analysis. The new dissimilarity emphasizes the importance of small eigenvalues which are known to carry the main information on graphs. General properties of the dissimilarity are discussed. The dissimilarity provides an efficient and intuitive tool for graph analysis.

1. Introduction

Graphs have been a topic of much interest due to many emerging data analysis applications having graph representation. For that reason, graph data analysis has been one of the focuses. To name a few, such problems exist in Systems Biology [7], Chemoinformatics [11] and web data. One of the tasks of data analysis is to define similarities or distances among structured objects, comparing graphs is a topic of much interest. The comparison is common in various scenarios such as similarity graph search, QSAR, or machine learning on graphs such as graph kernels [12].

Our target here is the problem of comparing different labeled graphs *sharing the same node set*. This is a special case of comparing graphs in general. While general graph comparison methods can be applied, there are methods that can only be applied in this particular problem setting. This problem

* Corresponding author. Tel.: +33 320434226.
E-mail addresses: nicolas.wicker@math.univ-lille1.fr (N. Wicker), canhhao@kuicr.kyoto-u.ac.jp (C.H. Nguyen), mami@kuicr.kyoto-u.ac.jp (H. Mamitsuka).

0024-3795/$ - see front matter
http://dx.doi.org/10.1016/j.laa.2012.10.021

arises in the case of comparing biological networks to see the similarities and differences between species, building phylogenetic trees [6].

The simplest solution is to consider the adjacency matrices as vectors and compare them directly. The most famous one is certainly the edit-distance [10,4], also named Levenshtein distance, which was initially developed to compare strings [8] and whose principle is simply to count the number of edges that are present in one graph and not in the other. Many other distance measures approximate edit-distances to account for its high computational complexity [13]. In this particular problem setting, edit-distance naturally becomes the number of different edges in the graphs. This means that all edges in the graph are considered of the same importance. This does not take into account global graph structures, which could be a problem in the case that global graph structures matter.

Many other measures are based on substructures such as maximal common subgraph [2,5]. This is based on the hypothesis that the semantics of the whole graph structures are based on the semantics of their subgraphs. The candidate subgraphs usually are walks, paths, frequent subgraphs [12]. In these methods, graph comparison bases solely on the existence of subgraphs. These measures fail to keep the graph structure as a whole and may not contain global structures for our interest. An attempt of using the global graph structures is to use graph spectra to reduce the problem of comparing graphs to the problem of comparing vectors [9].

In this work, we use spectral graph theory to compare graph in order to take into account *global graph structures*. We show a general framework and property of graph comparison using graph spectra. We show that some other similarity or dissimilarity measures are just special cases. A problem of the framework is that one needs to do a spectral transformation that gives high weights to the eigenvalues that are close to zero [3], and also ignores zero ones. Another problem is that graph comparison has to be invariant under different eigenspace bases, a problem of spectral representation. We propose a new dissimilarity measure that tackles these problems directly. Its advantages are shown on some canonical examples and as well as its properties.

2. Graph Laplacian-based graph comparison

We show a simple example in which a graph is considered as a matrix or vector in Fig. 1. It is noteworthy that this is equivalent to the edit-distance for our problem setting. We show that this distance is not adequate. The reason is that graphs have structures that are not easily seen in matrices or vectors. This motivates to use a representation that takes structure information into account. For that reason, we use the structure information contained in eigenvectors and eigenvalues of graph normalized Laplacians, henceforth simply called Laplacians.

In Fig. 1, we have five graphs G_1, G_2, G_3, G_4 and G_5. The difference between G_1 and G_2 is only one edge. There is also only one edge difference between G_1 and G_3 but G_3 is not connected. While considering graphs as matrices or vectors, the distance between G_1 and G_2 is the same as between G_1 and G_3. However, G_3 is totally different as it is not connected. Graph Laplacian can show this information in its eigenspectrum. G_4 and G_5 are two extremal graphs, namely the totally disconnected graph composed of 18 vertices and K_{18} the complete graph of size 18.

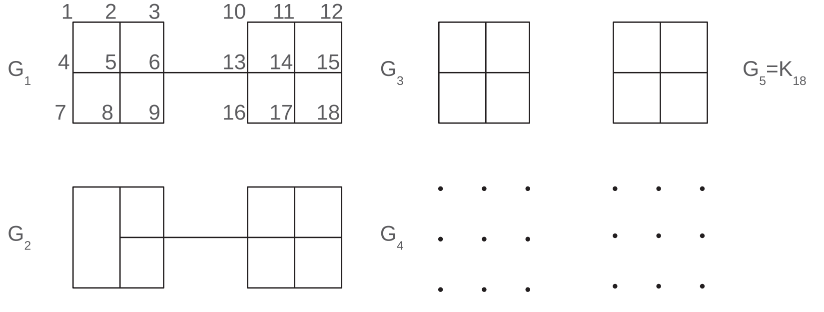

Fig. 1. Three graphs compared with our similarity.

Table 1
Correlation-based similarity.

	G_1	G_2	G_3	G_4	G_5
G_1	1	0.996	0.997	NaN	0.888
G_2	0.996	1	0.993	NaN	0.882
G_3	0.997	0.993	1	NaN	0.883
G_4	NaN	NaN	NaN	NaN	NaN
G_5	0.888	0.882	0.883	NaN	1

Table 2
Bregman divergences.

G_1	G_2	G_3	G_4	G_5
0	0.210	0.123	24.127	5.118
0.210	0	0.33	24.414	5.412
0.123	0.33	0	24.375	5.364
24.127	24.414	24.375	0	19.059
5.118	5.412	5.364	19.059	0

2.1. General graph Laplacian-based graph comparison

Let us consider two graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ which share the same set of vertices. Their Laplacians $\mathcal{L}_1$ and $\mathcal{L}_2$ have eigenvalues in increasing order $\lambda_1, \ldots, \lambda_n$ and $\mu_1, \ldots, \mu_n$ respectively. The corresponding eigenvectors are denoted $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$. We suppose that they are orthonormal, which is always possible as the Laplacians are symmetric.

We propose a general framework to compare graphs in the following form:

$$F(G_1, G_2) = \sum_{i,j} f(\lambda_i, \mu_j) |\langle u_i, v_j \rangle|^k, \tag{1}$$

for any $k \in N, k > 0$. The function $f : (\mathbb{R}, \mathbb{R}) \to \mathbb{R}$ is a comparison between eigenvalues. The measure F is a similarity or dissimilarity function according to f.

Special realizations of this measure are:

1. Correlation-based similarity. A natural similarity where unit length normalization of eigenvalues and dot product are used in f.

$$C(G_1, G_2) = \frac{1}{\sqrt{\sum_i \lambda_i^2 \sum_i \mu_i^2}} \sum_{i,j} \lambda_i \mu_j \langle u_i, v_j \rangle^2.$$

2. Bregman divergence (dissimilarity) with squared norm [1].

$$B(G_1, G_2) = \sum_{i,j} (\lambda_i - \mu_j)^2 \langle u_i, v_j \rangle^2.$$

3. New dissimilarity measure. We propose a new dissimilarity measure between the two graphs as follows:

$$D(G_1, G_2) = \sum_{i,j} \frac{(\lambda_i - \mu_j)^2}{\lambda_i + \mu_j} \langle u_i, v_j \rangle^2. \tag{2}$$

Values are given for all graph pairs of Fig. 1 in Tables 1–3. Interestingly, $D(G_1, G_2) = 0.122$ and $D(G_1, G_3) = 0.124$ showing that G_1 is closer to G_2 than to G_3. Thus, the lack of connectivity has a cost in the distance, due to the weighting by the eigenvalue inverses. This must be compared to the corresponding values obtained for the Bregman divergence, G_1 is then closer to G_3 than to G_2 which

Table 3
New dissimilarity measure.

G_1	G_2	G_3	G_4	G_5
0	0.122	0.124	18	2.876
0.122	0	0.244	18	3.057
0.124	0.244	0	18	3.078
18	18	18	0	18
2.876	3.057	3.078	18	0

is not a desirable result. On the contrary, the correlation based similarity has the same feature as the proposed dissimilarity, making G_1 closer to G_2 than to G_3. The drawback of this similarity is that when all eigenvalues are equal to 0 or close to it the normalization factor cannot be computed anymore. Let us remark at this point that our dissimilarity measure can also be naturally extended by continuity when eigenvalues are equal to 0. Indeed, if we consider two small eigenvalues x and y. If they are equal, $(x-y)^2/(x+y) = 0$, otherwise let us suppose without loss of generality that $x > y$, then $(x-y)^2/(x+y) < (x-y)^2/(x-y) = x-y$ which tends to 0 when x tends to 0. Considering these results, we have focused our study on the new dissimilarity d. It is indeed a dissimilarity and not a distance, it is sufficient to notice in the previous example that $D(G_3, G_1) + D(G_1, G_5) < D(G_3, G_5)$ to make it clear.

2.2. Invariance property

We show that among all those possible graph comparisons in (1), only for $k = 2$, the comparisons are invariant under different eigenvector bases of the graph Laplacian. Since the eigenvector bases are not supposed to be unique, we mean that only $k = 2$ should be used for all these graph comparisons.

Theorem 1. *The similarity/dissimilarity measures*

$$F(G_1, G_2) = \sum_{i,j} f(\lambda_i, \mu_j) |\langle u_i, v_j \rangle|^k$$

are invariant to the choice of eigenspace bases if and only if $k = 2$.

Proof. We want to prove that the measure F is invariant regardless of the choices of eigenvector bases. Since choices only happen for eigenvectors of the same eigenvalues, it is sufficient to prove that F is invariant if and only if $k = 2$.

Without loss of generality, if we suppose that $v_1, \ldots, v_l$ and $v_1^{'}, \ldots, v_l^{'}$ are two eigenvector bases corresponding to one eigenvalue μ of $\mathcal{L}_2$. The invariance of F can be boiled down to its invariance in the two bases. The necessary and sufficient condition for F to be invariant for any f is that for any unit vector u

$$\sum_{i=1}^{l} f(\lambda, \mu) |\langle u, v_i \rangle|^k = \sum_{i=1}^{l} f(\lambda, \mu) |\langle u, v_i^{'} \rangle|^k. \tag{3}$$

This is equivalent to:

$$\sum_{i=1}^{l} |\langle u, v_i \rangle|^k = \sum_{i=1}^{l} |\langle u, v_i^{'} \rangle|^k.$$

It is easy to see that for $k = 2$, this quantity is the length of the projection of u in the eigensubspace of μ.

Now we prove the other way around that, for $k \neq 2$, the equality in (3) does not hold in general. We construct a general counterexample as follow. Since the sets $\{v_1, \ldots, v_l\}$ and $\{v_1^{'}, \ldots, v_l^{'}\}$ are distinct, we can always choose a vector u in the former set not present in the latter one. Then,

$$\sum_{i=1}^{l} |\langle u, v_i \rangle|^k = 1,$$

because u is in the set. On the other hand,

$$\sum_{i=1}^{l} |\langle u, v_i^{'} \rangle|^k \quad > \quad \sum_{i=1}^{l} \langle u, v_i^{'} \rangle^2 = 1$$

for $k < 2$, and

$$\sum_{i=1}^{l} |\langle u, v_i^{'} \rangle|^k \quad < \quad \sum_{i=1}^{l} \langle u, v_i^{'} \rangle^2 = 1$$

for $k > 2$. Therefore, in general, the equality in (3) does not hold for $k \neq 2$. □

Corollary 1. *The dissimilarity function we proposed is invariant under any choice of eigenspace base.*

This desirable property has been proved in Theorem 1, remarking that $|\langle u, v \rangle^2| = \langle u, v \rangle^2$.

3. Properties of the new dissimilarity measure

We can show some properties of the new dissimilarity measure. In particular, it behaves well with regard to graph connectivity.

Theorem 2. *Dissimilarity $D(G_1, G_2) = 0$ implies that the Laplacians $\mathcal{L}_1$ and $\mathcal{L}_2$ eigendecompositions of G_1 and G_2 are equal.*

Proof. First we prove that $\forall i \in 1, \ldots, n$, $\lambda_i = \mu_i$. Let us consider that there is an eigenvalue λ which has multiplicity m_1 in $\mathcal{L}_1$ and m_2 in $\mathcal{L}_2$ with $m_1 \neq m_2$. Then, the subspace $E_\lambda^\perp$ orthogonal to the eigenspace E_λ of $\mathcal{L}_1$ for eigenvalue λ has dimension $n - m_1$. As $D(G_1, G_2) = 0$ the eigenvectors corresponding to λ in $\mathcal{L}_2$ are orthogonal to $E_\lambda^\perp$, so that $E_\lambda^\perp$ has dimension $n - m_2$ leading to a contradiction.

Now, we can prove that the eigenspaces are the same. If we take any eigenvalue λ, for $\mathcal{L}_1$, obviously, E_λ is orthogonal to all other eigenspaces. As for $\mathcal{L}_2$, its eigenspace F_λ for eigenvalue λ is also orthogonal to $E_\lambda^\perp$ as $D(G_1, G_2) = 0$, so that $F_\lambda = E_\lambda$. This is true for every eigenvalue λ and so the proof is completed. □

An important consequence of this theorem is that if $D(G_1, G_2) = 0$, the Laplacians and hence the graphs are the same.

4. Experiments

We conducted some experiments to demonstrate the properties of our dissimilarity measure in comparison with the correlation-based similarity, the Bregman divergence and the edit distance. The experimental setting was as follows. We started with a canonical graph G consisting of 9 disconnected

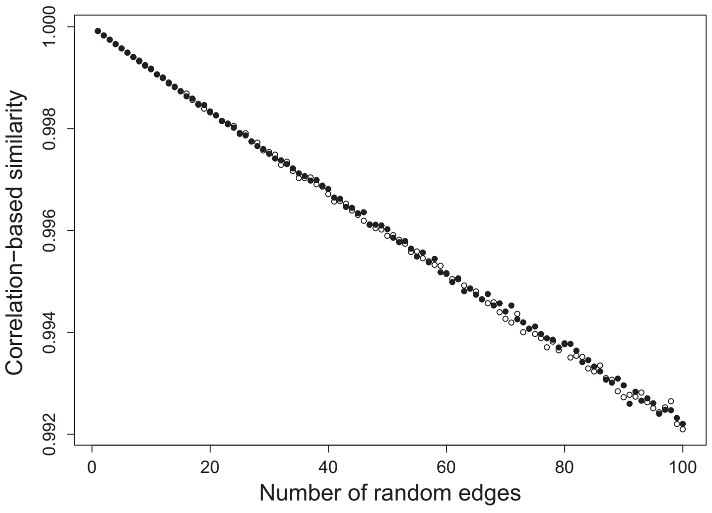

Fig. 2. Results for the correlation-based similarity. Horizontal axis shows the indices of graphs W_i and B_i and vertical axis shows the correlation-based similarity: $C(G, W_i)$ (bullets) and $C(G, B_i)$ (circles).

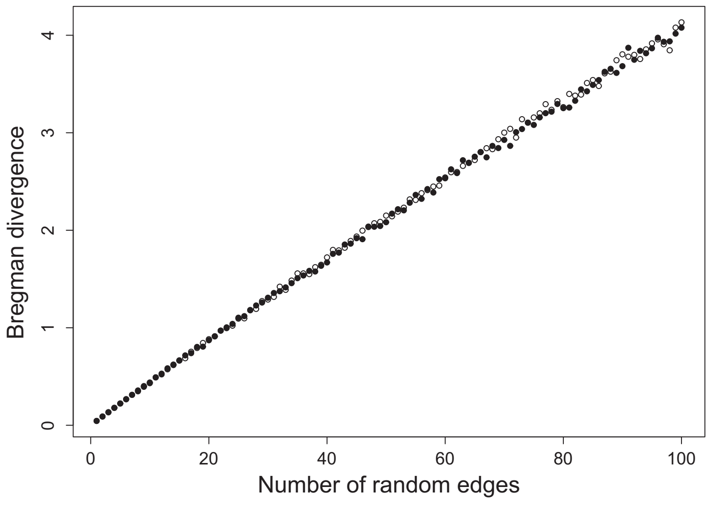

Fig. 3. Results for Bregman divergence. Horizontal axis shows the indices of graphs W_i and B_i and vertical axis shows the Bregman divergence: $B(G, W_i)$ (bullets) and $B(G, B_i)$ (circles).

subgraphs. Each subgraph was a 3-regular graph with 25 nodes. We generated 100 graphs, denoted $B_i, \forall i = 1 \cdots 100$, by adding randomly i edges connecting two nodes belonging to 2 different subgraphs of the 9 subgraphs (*between subgraph edges*). We also generated 100 other graphs, denoted $W_i, \forall i = 1 \cdots 100$, by adding randomly i edges connecting any two nodes in the same subgraph of the 9 subgraphs of G (*within subgraph edges*). It was our idea to generate the B_i graphs so that by adding edges between subgraphs, the connectivity of the graphs would change more than in the case of adding edges within subgraphs in the W_i graphs (the former case connects disconnected subgraphs while the latter does not). We wished dissimilarity measures reflect that graphs B_i are more dissimilar to G than its counterpart, W_i to G.

Experimental results are shown in Figs. 2, 3 and 4 for the correlation-based similarity, the Bregman divergence and the new dissimilarity measure. The following are observed.

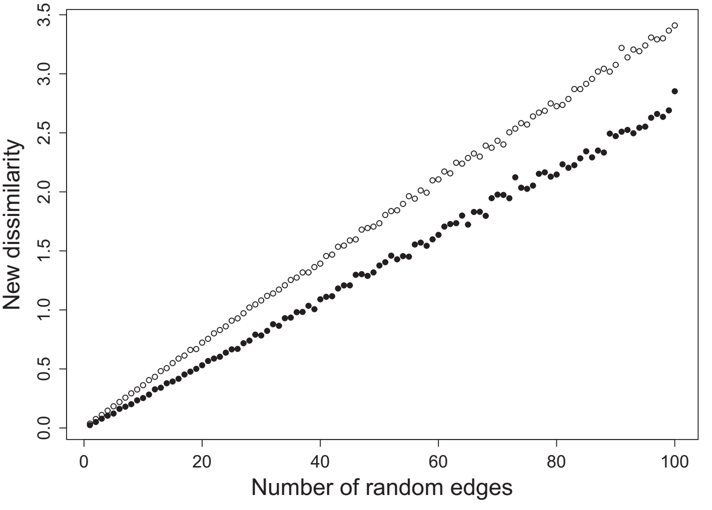

Fig. 4. Results for the new dissimilarity measure. Horizontal axis shows the indices of graphs W_i and B_i and vertical axis shows the dissimilarity measure: $D(G, W_i)$ (bullets) and $D(G, B_i)$ (circles).

1. The edit-distance, in our problem setting being the number of different edges, is the same for the B_i and W_i graph pairs, so graphs are not shown for it.
2. The correlation-based similarity and the Bregman divergence in Figs. 2 and 3 respectively, do not show much difference between the B_i and W_i graph pairs.
3. The new dissimilarity measure, as in Fig. 4, shows for the B_i and W_i graph pairs that graph B_i are more dissimilar than W_i, as we wish for in our experimental setup.

We conclude that, our proposed new dissimilarity measure is able to distinguish the differences of graphs with different graph connectivities while the others are not.

5. Conclusion

We have presented a framework for comparing graphs with the same node set, taking into account global graph structures. Properties of the framework are shown as well as special cases, including a new graph dissimilarity that is straight-forward to compute and, at the same time, has some nice properties. First, it is invariant under different graph eigenspace bases, making it independent of the eigendecomposition process. Second, a zero dissimilarity actually indicates that the two graphs are equal. Then, we have specifically aimed at having a dissimilarity giving more weight to eigenspaces with small eigenvalues. The measure proves to be able to take into account global structures in the toy example and the experiments.

References

[1] A. Banerjee, S. Merugu, I.S. Dhillon, J. Ghosh, Clustering with Bregman divergences, J. Mach. Learn. Res. 6 (2005) 1705–1749.
[2] H. Bunke, K. Shearer, A graph distance metric based on the maximal common subgraph, Pattern Recognit. Lett. 19 (3–4) (1998) 255–259.
[3] F.R.K. Chung, Spectral Graph Theory, American Mathematical Society, 1997.
[4] M.A. Eshera, K.S. Fu, A graph distance measure for image analysis, IEEE Trans. Syst. Man Cybern. 14 (1984) 398–407.
[5] M.-L. Fernández, G. Valiente, A graph distance metric combining maximum common subgraph and minimum common supergraph, Pattern Recognit. Lett. 22 (6–7) (2001) 753–758.
[6] M. Heymans, A.K. Singh, Deriving phylogenetic tree from the similarity analysis of metabolic pathways, Bioinformatics 26 (17) (2003) 138–146.
[7] H. Hu, X. Yan, Y. Huang, J. Han, X.J. Zhou, Mining coherent dense subgraphs across massive biological networks for functional discovery, Bioinformatics 21 (1) (2005) 213–221.

[8] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Sov. Phys. Dokl. 10 (1966) 707–710.
[9] B. Luo, R.C. Wilson, E.R. Hancock, Spectral embedding of graphs, Pattern Recognit. 36 (2003) 2213–2230.
[10] A. Sanfeliu, K.S. Fu, A distance measure between attributed relational graphs for pattern recognition, IEEE Trans. Syst. Man Cybern. 13 (1983) 353–362.
[11] N. Trinajstic, J.V. Knop, W.R. Muller, K. Syzmanski, S. Nikolic, Computational Chemical Graph Theory: Characterization, Enumeration and Generation of Chemical Structures by Computer Methods, Prentice Hall, 1991.
[12] S.V.B. Vishwanathan, N.N. Schraudolph, R. Kondor, K.M. Borgwardt, Graph kernels, J. Mach. Learn. Res. 11 (2010) 1201–1242.
[13] Z. Zeng, A.K.H. Tung, J. Wang, J. Feng, L. Zhou, Comparing stars: on approximating graph edit distance, Proc. VLDB Endowment 2 (1) (2009) 25–36.