Logarithmic proximity measures outperform plain ones in graph nodes clustering

Vladimir Ivashkin and Pavel Chebotarev

Institute of Control Sciences of the Russian Academy of Sciences and Moscow Institute of Physics and Technology

Abstract

We consider a number of graph kernels and proximity measures: commute time kernel, regularized Laplacian kernel, heat kernel, communicability, etc., and the corresponding distances as applied to clustering nodes in random graphs. The model of generating graphs involves edge probabilities for the pairs of nodes that belong to the same class or different classes. It turns out that in most cases, logarithmic measures (i.e., measures resulting after taking logarithm of the proximities) perform much better while distinguishing classes than the “plain” measures. A direct comparison of inter-class and intra-class distances confirms this conclusion. A possible explanation of this fact is that most kernels have a multiplicative nature, while the nature of distances used in cluster algorithms is an additive one (cf. the triangle inequality). The logarithmic transformation is just a tool to transform one nature to another. Moreover, some distances corresponding to the logarithmic measures possess a meaningful cutpoint additivity property [Che13]. In our experiments, the leader is the so-called logarithmic communicability measure, which distinctly outperforms the “plain” communicability.

1 Introduction

In this paper, we consider a number of graph kernels and proximity measures and the corresponding distances as applied to clustering nodes in random graphs. The measures are: the commute time kernel, the regularized Laplacian kernel, the heat kernel, communicability, and some others. The model of generating graphs involves edge probabilities for the pairs of nodes that belong to the same class or different classes. The main result is that in most cases, logarithmic measures (i.e., measures resulting after taking logarithm of the proximities) perform much better while distinguishing classes than the “plain” measures. A direct comparison of inter-class and intra-class distances confirms this conclusion.

2 Measures

In this study, we consider the following\(^1\) graph measures. Recall that if a graph proximity measure satisfies the triangle inequality for proximities \(p(x, y) + p(x, z) - p(y, z) \leq p(x, x)\) for all nodes \(x, y, z \in V(G)\), then the function \(d(x, y) = p(x, x) + p(y, y) - p(x, y) - p(y, x)\) satisfies the ordinary triangle inequality [CS98a].

2.1 The Shortest path and Commute time distances

- The Shortest Path distance \(d^s(i, j)\) on a graph \(G\) is the length of a shortest path between \(i\) and \(j\) in \(G\) [BH90].

---

\(^1\)On various graph kernels, see, e.g., [FFY+12].
• The **Commute Time** distance \(d^c(i,j)\) is an average length of random walks from one vertex to the other and back. It is related to the Commute-time kernel \([SFYD04]\) \(L^+\), the generalized inverse of the Laplacian matrix of \(G\).

• The **Resistance** distance \([Sha67,GG87,KR93]\) \(d^r(i,j)\) is the effective resistance between \(i\) and \(j\) in the electrical network corresponding to \(G\).

The **Resistance** distance is well known \([NW59,GJ74,CRR89]\) to be proportional to the **Commute Time** distance in the corresponding Markov chain.

As we mainly study parametric families of graph measures, for comparability, the parametric family \((1 - \lambda)D^s + \lambda D^r\) with \(\lambda \in [0, 1]\), i.e., the convex combination of the **Shortest Path** distance and the **Resistance** distance will be considered. It will be denoted by SP–CT.

### 2.2 The plain Walk, Forest, Communicability and Heat proximities

• **plain Walk** (Von Neumann diffusion kernel) \(H^{\text{pWalk}}_t = (I - tA)^{-1}, 0 < t < \rho^{-1}\) (\(\rho\) is the spectral radius of \(A\), the adjacency matrix of the graph \(G\)) \([CS98b,KCST02]\).

• **Forest** (Regularized Laplacian kernel): \(H^{\text{For}}_t = (I + tL)^{-1}, t > 0\), where \(L\) is the Laplacian matrix of \(G\) \([CS97,CS02,SK03]\).

• **Communicability** (Exponential diffusion kernel): \(H^{\text{Comm}}_t = \exp(tA), t > 0\) \([KL02,Est12]\).

• **Heat kernel** (Laplacian exponential diffusion kernel): \(H^{\text{Heat}}_t = \exp(-tL), t > 0\) \([CY98,KL02]\).

### 2.3 Logarithmic measures \([Che11b]\): Walk, Forest, Communicability, and Heat

• **Walk** (logarithmic): \(H^{\text{Walk}}_t = \ln H^{\text{pWalk}}_t, 0 < t < \rho^{-1}\), \(\ln\) is elementwise \(\ln\) \([Che12]\).

• **logarithmic Forest**: \(H^{\text{logFor}}_t = \ln H^{\text{For}}_t, t > 0\) \([Che11a]\).

• **logarithmic Communicability**: \(H^{\text{logComm}}_t = \ln H^{\text{Comm}}_t, t > 0\).

• **logarithmic Heat**: \(H^{\text{logHeat}}_t = \ln H^{\text{Heat}}_t, t > 0\).

### 2.4 Randomized Shortest Path and Free Energy \([KSS14]\)

\[
P^{\text{pref}} = D^{-1}A, \quad D = \text{Diag}(Ac);
\]
\[
W = P^{\text{pref}} \circ \exp(-\beta C); \quad \circ \text{ is elementwise product,}
\]
\(C\) is the matrix of the Shortest Path distances;
\[
Z = (I - W)^{-1}.
\]

• **Randomized Shortest Path**:

\[
S = (Z(C \circ W)Z) / Z; \quad \div \text{ is elementwise division;}
\]
\[
\bar{C} = S - e(d_S)^T; \quad d_S = \text{diag}(S);
\]
\[
\Delta_{\text{RSP}} = (\bar{C} + \bar{C}^T)/2.
\]
- Helmholtz Free Energy distance:

\[ Z^h = ZD^{-1}_h; \quad D_h = \text{Diag}(Z); \]
\[ \Phi = -1/\beta \log(Z^h); \]
\[ \Delta_{FE} = (\Phi + \Phi^T)/2. \]

For comparability, all family parameters are adapted to the [0, 1] segment by a linear transformation or the \( t/(t + 1) \) transformation.

### 3 Logarithmic vs. plain measures

\( G(N, (m)p_{in}, p_{out}) \) is the model of generating graphs on \( N \) nodes divided into \( m \) classes of equal (or almost equal) size, with \( p_{in} \) and \( p_{out} \) being the probability of \( (i, j) \in E(G) \) for \( i \) and \( j \) belonging to the same class and different classes, respectively.

The curves below present the Rate index\(^2\) (averaged over 200 random graphs) for clustering with Ward method.

![Graphs showing comparisons](image)

(a) plain Walk and Walk

(b) Forest and logarithmic Forest

(c) Communicability and logarithmic Communicability

(d) Heat and logarithmic Heat

Figure 1: Logarithmic vs. plain measures for \( G(100, (2)0.2, 0.05) \)

\(^2\)Also referred to as the Rand index.
It can be seen that in almost all cases, logarithmic measures majorize the ordinary ones. The only exception is the case of Walk measures for graphs on 200 nodes.

4 Competition by Copeland’s score

4.1 Approach [KSS14]

• The competition of measure families is based on paired comparisons.

• Every time the best Rate index (RI) of a measure family \( F_1 \) is higher on a test graph than that of some other measure family \( F_2 \), we add +1 to the score of \( F_1 \) and −1 to the score of \( F_2 \).

4.2 The competition results

The competition has been performed on random graphs generated with the \( G(n,p_{in};p_{out}) \) model and the following parameters: \( n \in \{100, 150\} \), number of clusters \( \in \{2, 4\} \), \( p_{in} = 0.3 \), \( p_{out} = 0.3 \). 

4
$p_{\text{out}} \in \{0.1, 0.15\}$. For every combination of parameters, we generated 5 graphs and for each of them we computed the best RIs the measure families reached.

| Nodes  | 100  | 100  | 100  | 100  | 150  | 150  | 150  | 150  |
|--------|------|------|------|------|------|------|------|------|
| Clusters | 2  | 2  | 4  | 4  | 2  | 2  | 4  | 4  |
| $p_{\text{out}}$ | 0.1 | 0.15 | 0.1 | 0.15 | 0.1 | 0.15 | 0.1 | 0.15 |
| logComm | 36 | 48 | 36 | 6 | 32 | 50 | 50 | 48 | 306 |
| logFor | 21 | 13 | 36 | 22 | -4 | 16 | 24 | 26 | 154 |
| logHeat | 26 | 21 | 24 | 8 | 28 | 19 | 24 | -2 | 148 |
| FE | 9 | 21 | 26 | 18 | 1 | 16 | 4 | 32 | 127 |
| Walk | -1 | 3 | 16 | 12 | -23 | 7 | -6 | 18 | 26 |
| Comm | -5 | 2 | -3 | -16 | 24 | 4 | 8 | 0 | 14 |
| pWalk | -2 | -8 | -11 | 18 | -6 | 10 | 6 | 0 | 7 |
| RSP | -13 | 20 | -8 | 0 | 7 | -15 | 10 | -15 | -14 |
| Heat | 18 | -39 | -40 | -40 | 23 | -23 | -40 | -45 | -186 |
| SP-CT | -41 | -32 | -26 | 22 | -38 | -34 | -32 | -17 | -198 |
| For | -48 | -49 | -50 | -50 | -44 | -50 | -48 | -45 | -384 |

Static parameters: $p_{\text{in}} = 0.3$; 5 graphs for each competition.

4.3 A similar competition for 90th percentiles

When we are looking for the best parameter of a measure family, we compute RI on a grid of that parameter. In the above competition, we only compared the highest RI values. Now consider the set of RI values some measure family provides as a sample and find its 90th percentile. These percentiles become the participants in a new tournament. The motivation behind this approach is to take into account the robustness of each family.

The results of this competition are given below.

| Nodes  | 100  | 100  | 100  | 100  | 150  | 150  | 150  | 150  |
|--------|------|------|------|------|------|------|------|------|
| Clusters | 2  | 2  | 4  | 4  | 2  | 2  | 4  | 4  |
| $p_{\text{out}}$ | 0.1 | 0.15 | 0.1 | 0.15 | 0.1 | 0.15 | 0.1 | 0.15 |
| logComm | 44 | 50 | 36 | 22 | 44 | 48 | 50 | 50 | 344 |
| logFor | 6 | 25 | 36 | 24 | 5 | 19 | 24 | 6 | 145 |
| Walk | 11 | 21 | 24 | 24 | -11 | 10 | 22 | 24 | 125 |
| FE | -2 | 15 | 14 | 24 | 1 | 9 | 20 | 22 | 103 |
| logHeat | 34 | -6 | 26 | -18 | 37 | 15 | 20 | -24 | 84 |
| pWalk | -7 | 25 | 4 | 15 | -6 | 18 | 10 | 22 | 81 |
| Comm | 0 | -10 | -6 | -23 | 0 | 7 | -12 | -14 | -58 |
| RSP | -28 | -10 | -22 | -4 | -28 | -20 | -20 | 0 | -132 |
| SP-CT | -41 | -20 | -22 | 26 | -46 | -32 | -24 | 4 | -155 |
| Heat | 27 | -40 | -40 | -40 | 37 | -24 | -40 | -40 | -160 |
| For | -44 | -50 | -50 | -50 | -33 | -50 | -50 | -50 | -377 |

We see that the order of families provided by the second competition has a number of differences from that given by the first one.

5 Reject curves

5.1 Definition

The ROC curve (also referred to as the Reject curve) in this case can be defined as follows.
- Create a grid of thresholds of the distance matrix values from minimum to maximum.

- For each threshold count a number of inter-cluster and intra-cluster distances which are less or equal than threshold.

- The Reject curve is the dependency of the “percentage of intra-cluster distances under the threshold” upon the “percentage of inter-cluster ones”.

A better measure is characterized by a curve that goes higher or, at least, has a larger area under curve.

5.2 Results

The optimum values of the family parameters (adjusted to the [0, 1] segment) w.r.t. the RI in the Ward method clustering for several $G(N, (m)p_{in}, p_{out})$ models are presented in the following table. The optimum is chosen on the grid of 50 parameter values. The second number in each cell of the table is the RI averaged over 200 random graphs.

| Kernel   | $G(100, (2)^{0.3}, 0.02)$ | $G(100, (2)^{0.3}, 0.05)$ | $G(100, (2)^{0.3}, 0.1)$ | $G(100, (2)^{0.3}, 0.15)$ |
|----------|---------------------------|---------------------------|--------------------------|---------------------------|
| pWalk K  | 0.8913, 0.9784            | 0.9348, 0.9708            | 0.7826, 0.8925           | 0.7174, 0.7015            |
| Walk K   | 0.9978, 0.9790            | 0.8261, 0.9690            | 0.6957, 0.8921           | 0.7174, 0.7099            |
| For K    | 1.0000, 0.9790            | 1.0000, 0.9640            | 0.9348, 0.7136           | 0.0000, 0.4957            |
| logFor K | 0.6304, 0.9794            | 0.6522, 0.9711            | 0.1957, 0.9009           | 0.1957, 0.7212            |
| Comm K   | 0.3261, 0.9798            | 0.2609, 0.9776            | 0.1957, 0.9170           | 0.1522, 0.7200            |
| logComm K| 0.3606, 0.9800            | 0.4130, 0.9793            | 0.3913, 0.9579           | 0.6087, 0.8323            |
| Heat K   | 0.8478, 0.9800            | 0.5652, 0.9794            | 0.4783, 0.9390           | 0.5652, 0.6725            |
| logHeat K| 0.8696, 0.9800            | 0.5652, 0.9794            | 0.4565, 0.9394           | 0.1522, 0.7194            |
| RSP K    | 0.9783, 0.9795            | 0.9783, 0.9714            | 0.9783, 0.8986           | 0.9783, 0.7138            |
| FE K     | 0.9783, 0.9791            | 0.9783, 0.9702            | 0.9130, 0.8988           | 0.9348, 0.7177            |
| SP-CT K  | 1.0000, 0.9790            | 1.0000, 0.9640            | 0.9783, 0.7980           | 0.9783, 0.6327            |

The reject curves for $G(100, (2)^{0.3}, 0.1)$ and optimum values of the family parameters (w.r.t. the Ward clustering RI) are given below. The pivot points of the curves are shown separately for each of 200 random graphs.
Figure 4: Reject curves for graph measures

The $\varepsilon$-like bend of several curves appears because the corresponding measures strongly depend on the Shortest path (SP) distance between nodes. In our experiments, this distance only takes few small values.

Finally, we show the reject curves averaged over 200 random graphs. The curves for the four leading families are duplicated separately.

Figure 5: Average reject curves

6 Conclusion

The main result of our study is that in most cases, logarithmic measures (i.e., measures resulting after taking logarithm of the proximities) perform much better while distinguishing classes than the “plain” measures. A direct comparison of inter-class and intra-class distances confirms this conclusion.

A plausible explanation of the superiority of logarithmic measures is that most kernels and proximity measures under study have a multiplicative nature, while the nature of distances used in cluster algorithms is an additive one (cf. the triangle inequality). The logarithmic transformation is just a tool to transform one nature to another. Moreover, some distances corresponding to the logarithmic measures possess a meaningful cutpoint additivity property.

In our experiments, the four leading measures are: logarithmic Communicability, logarithmic Forest, logarithmic Heat, and Free Energy. The latter can also be considered as a kind of logarithmic measure: see the expression for $\Phi$ in its definition in Subsection 2.4. The excellence of logarithmic Communicability over all other measures under study in the present context is undoubted.
References

[BH90] Buckley, F.; Harary, F.: Distance in Graphs. Redwood City, CA: Addison-Wesley, 1990

[Che11a] Chebotarev, P.: A Class of graph-geodetic distances generalizing the shortest-path and the resistance distances. In: Discrete Applied Mathematics 159 (2011), Nr. 5, S. 295–302

[Che11b] Chebotarev, P.: The graph bottleneck identity. In: Advances in Applied Mathematics 47 (2011), Nr. 3, S. 403–413

[Che12] Chebotarev, P.: The walk distances in graphs. In: Discrete Applied Mathematics 160 (2012), Nr. 10–11, S. 1484–1500

[Che13] Chebotarev, P.: Studying new classes of graph metrics. In: Nielsen, F. (Hrsg.); Barbaresco, F. (Hrsg.): Proceedings of the SEE Conference “Geometric Science of Information” (GSI-2013). Berlin: Springer, 2013 (Lecture Notes in Computer Science, LNCS 8085), S. 207–214

[CRR+89] Chandra, A. K.; Raghavan, P.; Ruzzo, W. L.; Smolensky, R.; Tiwari, P.: The electrical resistance of a graph captures its commute and cover times. In: Proc. 21st Annual ACM Symp. on Theory of Computing. Seattle: ACM Press, 1989, S. 574–586

[CS97] Chebotarev, P. Y.; Shamis, E. V.: The Matrix-Forest Theorem and Measuring Relations in Small Social Groups. In: Automation and Remote Control 58 (1997), Nr. 9, S. 1505–1514

[CS98a] Chebotarev, P. Y.; Shamis, E. V.: On a Duality Between Metrics and $\Sigma$-proximities. In: Automation and Remote Control 59 (1998), Nr. 4, S. 608–612

[CS98b] Chebotarev, P. Y.; Shamis, E. V.: On Proximity Measures for Graph Vertices. In: Automation and Remote Control 59 (1998), Nr. 10, S. 1443–1459

[CS02] Chebotarev, Pavel; Shamis, Elena: The Forest Metrics for Graph Vertices. In: Electronic Notes in Discrete Mathematics 11 (2002), S. 98–107

[CY98] Chung, Fan; Yau, Shing-Tung: Coverings, heat kernels and spanning trees. In: Journal of Combinatorics 6 (1998), S. 163–184

[Est12] Estrada, Ernesto: The communicability distance in graphs. In: Linear Algebra and its Applications 436 (2012), Nr. 11, S. 4317–4328

[FFY+12] Fouss, F.; Francoisse, K.; Yen, L.; Pirotte, A.; Saerens, M.: An Experimental Investigation of Kernels on Graphs for Collaborative Recommendation and Semisupervised Classification. In: Neural Networks 31 (2012), July, S. 53–72

[GG87] Gvishiani, A. D.; Gurvich, V. A.: Metric and ultrametric spaces of resistances. In: Russian Mathematical Surveys 42 (1987), Nr. 6, S. 235–236

[GJ74] Göbel, F.; Jagers, A. A.: Random walks on graphs. In: Stochastic Processes and their Applications 2 (1974), Nr. 4, S. 311–336

[KCST02] Kandola, Jaz; Cristianini, Nello; Shawe-Taylor, John S.: Learning semantic similarity. In: Advances in neural information processing systems, 2002, S. 657–664
| Reference | Authors | Title | Venue | Pages |
|-----------|---------|-------|-------|-------|
| [KL02]    | Kondor, Risi I.; Lafferty, John | Diffusion kernels on graphs and other discrete structures | Proceedings of the 19th international conference on machine learning, 2002 | S. 315–322 |
| [KR93]    | Klein, D. J.; Randić, M. | Resistance distance | Journal of Mathematical Chemistry 12 (1993) | S. 81–95 |
| [KSS14]   | Kivimäki, Ilkka; Shimbo, Masashi; Saerens, Marco | Developments in the theory of randomized shortest paths with a comparison of graph node distances | Physica A: Statistical Mechanics and its Applications 393 (2014) | S. 600–616 |
| [NW59]    | Nash-Williams, C. St J. A. | Random walk and electric currents in networks | Mathematical Proceedings of the Cambridge Philosophical Society Bd. 55 Cambridge Univ Press, 1959 | S. 181–194 |
| [SFYD04]  | Saerens, Marco; Fouss, Francois; Yen, Luh; Dupont, Pierre | The principal components analysis of a graph, and its relationships to spectral clustering | Machine Learning: ECML 2004 | Springer, 2004, S. 371–383 |
| [Sha67]   | Sharpe, G. E. | Solution of the \((m + 1)\)-terminal resistive network problem by means of metric geometry | Proceedings of the First Asilomar Conference on Circuits and Systems | Pacific Grove, CA, November 1967, S. 319–328 |
| [SK03]    | Smola, Alex J.; Kondor, Risi I. | Kernels and Regularization of Graphs | Proceedings of the 16th Annual Conference on Learning Theory | 2003, S. 144–158 |