The Sum-Over-Paths Covariance: A novel covariance measure between nodes of a graph

Graph mining, kernel on a graph, correlation measure, semi-supervised classification.

Amin Mantrach AMANTRAC@ULB.AC.BE
Institut de Recherches Interdisciplinaires et de Développements en Intelligence Artificielle (IRIDIA-CoDE)
Université Libre de Bruxelles, B-1050 Brussels, Belgium

Marco Saerens MARCO.SAERENS@UCLOUVAIN.BE
Luh Yen LUH.YEN@UCLOUVAIN.BE
UCL Machine Learning Group (MLG), Louvain School of Management,
Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium

Abstract

This work introduces a link-based covariance measure between the nodes of a weighted, directed, graph where a cost is associated to each arc. To this end, a probability distribution on the (usually infinite) set of paths through the network is defined by minimizing the sum of the expected costs between all pairs of nodes while fixing the total relative entropy spread in the network. This results in a probability distribution on the set of paths such that long paths (with a high cost) occur with a low probability while short paths (with a low cost) occur with a high probability. The sum-over-paths (SoP) covariance measure is then computed according to this probability distribution: two nodes will be highly correlated if they often co-occur together on the same – preferably short – paths. The resulting covariance matrix between nodes (say n in total) is a Gram matrix and therefore defines a valid kernel matrix on the graph; it is obtained by inverting a $n \times n$ matrix. The proposed model could be used for various graph mining tasks such as computing betweenness centrality, semi-supervised classification, visualization, etc.

1. Introduction

Network and link analysis is an important, growing, field that has been the subject of much recent work in various fields of science: applied mathematics, computer science, social science, physics, pattern recognition. Within this context, one key issue is the proper definition of a similarity measure between the nodes of the network, taking both direct and indirect links into account (some examples are [2, 3, 5, 7]). This paper precisely proposes such a similarity measure, taking the form of a covariance matrix, by extending the framework developed in [6] in the context of routing. This quantity will be called the **sum-over-paths** (SoP) **covariance** and has a clear, intuitive, interpretation: two nodes are correlated if they often co-occur on the same – preferably short – paths (including cycles). To this end, a probability distribution on all possible paths through the network, inspired by [1], is defined by adopting a sum-over-paths statistical physics framework. This results in a probability distribution on the (usually infinite) set of paths such that long paths occur with a low probability while short paths occur with a high probability. The covariance measure between nodes is then computed according to this probability distribution. It characterizes the relations between the nodes and depends on a parameter, θ, controlling the entropy (exploration) spread in the network. Of course, the set of paths does not need to be enumerated; the covariance matrix is obtained by inverting a $n \times n$ matrix.

Preliminary work. Under review by the International Workshop on Mining and Learning with Graphs (MLG). Do not distribute.

2. The SoP covariance measure

Basic notations and definitions. Consider a weighted directed graph or network, G, with a set of n nodes V (or vertices) and a set of arcs E (or edges). The graph is supposed to be strongly connected. To each arc linking node k and node k', a number $c_{kk'} > 0$ is associated, representing the **immediate cost** of following this arc. The **cost matrix C** is the matrix containing the immediate costs $c_{kk'}$. In a first step, a **random walk** on this graph is defined in the usual way. The choice to follow an arc will be made according to transition probabilities rep-

resenting the probability of jumping from a node k to a node k' belonging to the set $S(k)$ of neighboring nodes (successors S). The transition probabilities defined on each node k will be denoted as $p_{kk'} = \mathrm{P}(k'|k)$ with $k' \in S(k)$. Furthermore, $\mathbf{P}$ will be the matrix containing the transition probabilities $p_{kk'}$ as elements. If there is no arc between k and k', we simply consider that $c_{kk'}$ takes a large value, denoted by ∞; in this case, the corresponding transition probability will be set to zero, $p_{kk'} = 0$. The *natural random walk* on the graph will be defined by $p^{\mathrm{ref}}_{kk'} = c^{-1}_{kk'} / \sum_{k'} c^{-1}_{kk'}$ and the corresponding transition-probabilities matrix $\mathbf{P}^{\mathrm{ref}}$. In other words, in this natural random walk, the random walker chooses to follow a link with a probability proportional to the inverse of the immediate cost, therefore favouring locally links having a low cost. These transition probabilities will be used as reference probabilities later; hence the superscript *ref*.

Definition of the probability distribution on the set of paths. Let us first consider two nodes, an initial node i and a destination node j. We define the (possibly infinite) set of paths (including cycles) connecting these two nodes as $\mathcal{R}_{ij} = \{\wp_{r^{ij}}\}$. Thus, $\wp_{r^{ij}}$ is path number r^{ij}, with path index r^{ij} ranging from 1 to ∞. Let us further define the set of all paths $\mathcal{R} = \bigcup_{ij} \mathcal{R}_{ij}$ and a probability distribution on this set $\mathcal{R}$ representing the probability $\mathrm{P}(\wp_{r^{ij}})$ of following the path numbered r^{ij}. The main idea will be to use the probability distribution $\mathrm{P}(\wp_{r^{ij}})$ *minimizing the expected cost-to-go among all the probability distributions having a fixed relative entropy with respect to the natural random walk on the graph.* This choice naturally defines a probability distribution on the set of paths such that long paths (with a high cost) occur with a low probability while short paths (with a low cost) occur with a high probability. Let us also denote as $E_{r^{ij}}$ the total cost associated to the path $\wp_{r^{ij}}$, referred to as the **energy** associated to that path. We assume that the total cost associated to a path is additive, i.e. $E(\wp_{r^{ij}}) = \sum_{t=1}^{t_f} c_{k_{t-1}k_t}$ where $k_0 = i$ is the initial node and $k_{t_f} = j$ is the destination node; t_f is the time (number of steps) needed to reach node j. Here, we assume that $\wp_{r^{ij}}$ is a valid path from the initial node to the destination node, that is, every $c_{k_{t-1}k_t} \neq \infty$ along that path.

We now have to find the path probabilities minimizing the sum of the expected energy for reaching node j when starting from i. In other works, we are seeking path probabilities, $\mathrm{P}(\wp_{r^{ij}})$, minimizing $\sum_{i,j=1}^{n} \sum_{r^{ij}=1}^{\infty} \mathrm{P}(\wp_{r^{ij}}) E(\wp_{r^{ij}})$ subject to the constraint $-\sum_{i,j=1}^{n} \sum_{r^{ij}=1}^{\infty} \mathrm{P}(\wp_{r^{ij}}) \ln(\mathrm{P}(\wp_{r^{ij}})/\mathrm{P}^{\mathrm{ref}}(\wp_{r^{ij}})) = J_0$ where $\mathrm{P}^{\mathrm{ref}}(\wp_{r^{ij}})$ represents the probability of following the path $\wp_{r^{ij}}$ when walking according to the natural random walk, i.e. when using transition probabilities $p^{\mathrm{ref}}_{kk'}$. Here, J_0 is provided a priori by the user, according to the desired degree of randomness he is willing to concede. By defining the Lagrange function

$$\begin{aligned}\mathcal{L} &= \sum_{i,j=1}^{n} \sum_{r^{ij}=1}^{\infty} \mathrm{P}(\wp_{r^{ij}}) E(\wp_{r^{ij}}) \\ &+ \lambda \left[\sum_{i,j=1}^{n} \sum_{r^{ij}=1}^{\infty} \mathrm{P}(\wp_{r^{ij}}) \ln \frac{\mathrm{P}(\wp_{r^{ij}})}{\mathrm{P}^{\mathrm{ref}}(\wp_{r^{ij}})} + J_0 \right] \\ &+ \mu \left[\sum_{i,j=1}^{n} \sum_{r^{ij}=1}^{\infty} \mathrm{P}(\wp_{r^{ij}}) - 1 \right], \end{aligned} \tag{1}$$

we obtain the following Boltzmann probability distribution

$$\mathrm{P}(\wp_{r^{ij}}) = \frac{\mathrm{P}^{\mathrm{ref}}(\wp_{r^{ij}}) \exp\left[-\theta E(\wp_{r^{ij}})\right]}{\sum_{i,j=1}^{n} \sum_{r^{ij}=1}^{\infty} \mathrm{P}^{\mathrm{ref}}(\wp_{r^{ij}}) \exp\left[-\theta E(\wp_{r^{ij}})\right]} \tag{2}$$

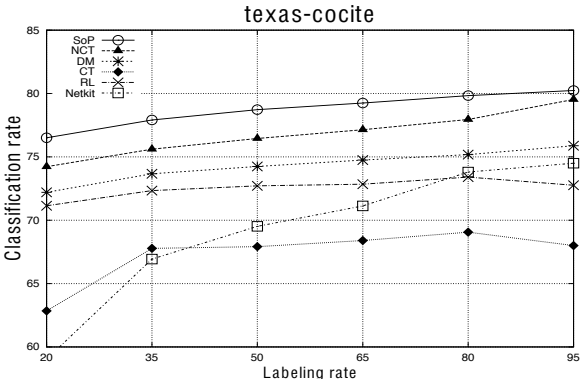
where $\theta = 1/\lambda$. Thus, as expected, short paths (having small $E(\wp_{r^{ij}})$) are favoured in that they have a large probability of being followed. When $\theta \to \infty$, only shortest paths are considered in $\mathcal{R}$ while when $\theta \to 0$ all paths corresponding to the natural random walk are taken into account, weighted by the likelihood of that path.

Definition of the covariance measure. We now show that the sum-over-paths covariance measure can be computed from a key quantity, defined as $\mathcal{Z} = \sum_{i,j=1}^{n} \sum_{r^{ij}=1}^{\infty} \mathrm{P}^{\mathrm{ref}}(\wp_{r^{ij}}) \exp\left[-\theta E(\wp_{r^{ij}})\right]$, which corresponds to the **partition function** in statistical physics. Indeed, by taking the second-order derivative of $\theta^{-2} \ln \mathcal{Z}$, we obtain the expected number of times the link $k \to k'$ and the link $l \to l'$ are traversed along the same path

$$\begin{aligned}\overline{\eta}(k,k';l,l') &= \frac{1}{\theta^2} \frac{\partial^2 (\ln \mathcal{Z})}{\partial c_{ll'} \partial c_{kk'}} \\ &= \sum_{i,j=1}^{n} \sum_{r^{ij}=1}^{\infty} \mathrm{P}(\wp_{r^{ij}}) \, \delta(r^{ij};k,k') \delta(r^{ij};l,l') \\ &- \left[\sum_{i,j=1}^{n} \sum_{r^{ij}=1}^{\infty} \mathrm{P}(\wp_{r^{ij}}) \, \delta(r^{ij};k,k') \right]^2 \end{aligned} \tag{3}$$

where $\delta(r^{ij};k,k')$ indicates the number of times the link $k \to k'$ is present in path number r^{ij}, and thus the number of times the link is traversed. This last quantity clearly corresponds to the **covariance** between link $k \to k'$ and link $l \to l'$.

Now, the SoP **covariance measure** between node k' and node l' is simply defined as $\mathrm{cov}(k',l') =$



$\sum_{k,l=1}^{n} \overline{\eta}(k,k';l,l')$ which corresponds to the main quantity of interest.

Computation of the partition function $\mathcal{Z}$. By using a trick introduced in [1], we now show how the partition function can be computed from the cost matrix (see [6] for details). Indeed, let us first build a new matrix, $\mathbf{W} = \mathbf{P}^{\text{ref}} \circ \exp\left[-\theta \mathbf{C}\right] = \exp\left[-\theta \mathbf{C} + \ln \mathbf{P}^{\text{ref}}\right]$, where $\mathbf{P}^{\text{ref}}$ is the transition-probabilities matrix containing the $p^{\text{ref}}_{kk'}$, and the logarithm/exponential functions are taken elementwise. Moreover, $\circ$ is the elementwise (Hadamard) matrix product. Furthermore, let $\mathbf{Z} = (\mathbf{I} - \mathbf{W})^{-1}$ and the element i,j of $\mathbf{Z}$ be z_{ij}. A few calculus shows that the partition function $\mathcal{Z}$ can be computed through $\mathcal{Z} = z_{\bullet\bullet} - n$ where $z_{\bullet\bullet} = \sum_{i,j=1}^{n} z_{ij}$ (see [6] for a similar derivation in another context).

Computation of the covariance measure. Taking the second-order derivative of $\theta^{-2} \ln \mathcal{Z}$ as in Equation (3) and rearranging the terms (the calculus is quite similar to the one appearing in [6]) yields the **covariance** between node k and node l:

$$\text{cov}(k,l) = \frac{1}{\mathcal{Z}} \Big\{ (z_{\bullet k} - 1) z_{k\bullet} \delta_{kl} + z_{k\bullet}(z_{\bullet l} - 1)(z_{lk} - \delta_{lk}) + z_{l\bullet}(z_{\bullet k} - 1)(z_{kl} - \delta_{kl}) - \frac{z_{k\bullet} z_{l\bullet}(z_{\bullet k} - 1)(z_{\bullet l} - 1)}{\mathcal{Z}} \Big\}$$

where δ_{kl} is the Kronecker delta and $z_{k\bullet} = \sum_{j=1}^{n} z_{kj}$. The matrix made of the z_{ij} is positive semi-definite and therefore defines a valid kernel on the graph. On the other hand, the **correlation** between nodes can be computed in the usual way by $\text{cor}(k,l) = \text{cov}(k,l)/\sqrt{\text{cov}(k,k)\text{cov}(l,l)}$.

3. Preliminary experiments

Preliminary experiments on semi-supervised classification of unlabeled nodes have been perfomed on the Texas Cocite dataset (described in [4]) and on other data sets not reported here. We compared the proposed SoP correlation kernel to (i) the normalized commute-time (NCT) kernel [7], (ii) the commute-time (CT) kernel [3], (iii) the diffusion map (DM) kernel [5, 3], (iv) the regularized Laplacian (RL) kernel [2, 3] and, as a baseline, (vi) the Netkit (Netkit) framework described in [4] with the default parameters present in the framework, which generally provide good results. A kernel alignment procedure is used in order to classify the unlabeled nodes, as described in [7], for each of the five kernels.

The classification accuracy is reported for increasing labeling rates, *i.e.* proportion of nodes for which the label is known. The labels of remaining nodes are removed and used as test. For each considered labeling rate, 100 random node label deletions (100 *runs*) were performed, on which performances are averaged. The hyper-parameters of each algorithm have been tuned within each run by using a nested 5-fold cross-validation. The classification rates for increasing proportions of labeled nodes are reported in the figure hereabove. By examining the results, we observe that the sum-over-paths kernel provides competitive results in comparison with the other standard kernels on a graph. This has been confirmed on various other data sets showing that the SoP and the NCT kernels are showing the best performances.

References

[1] T. Akamatsu. Cyclic flows, markov process and stochastic traffic assignment. *Transportation Research B*, 30(5):369–386, 1996.

[2] P. Chebotarev and E. Shamis. The matrix-forest theorem and measuring relations in small social groups. *Automation and Remote Control*, 58(9):1505–1514, 1997.

[3] F. Fouss, L. Yen, A. Pirotte, and M. Saerens. An experimental investigation of graph kernels on a collaborative recommendation task. *Proceedings of the 6th International Conference on Data Mining (ICDM 2006)*, pages 863–868, 2006.

[4] S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8:935–983, 2007.

[5] B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. *Advances in Neural Information Processing Systems 18*, pages 955–962, 2005.

[6] M. Saerens, Y. Achbany, F. Fouss, and L. Yen. Randomized shortest-path problems: Two seemingly unrelated problems. *Manuscript submitted for publication*, 2008.

[7] D. Zhou, J. Huang, and B. Scholkopf. Learning from labeled and unlabeled data on a directed graph. *Proceedings of the 22nd International Conference on Machine Learning*, pages 1041–1048, 2005.