Assortativity Patterns in Multidimensional Attributed Networks: a Statistical Approach

Konstantinos Pelechrinis*, Marios Kokkodis+ & Dong Wei*

*School of Information Sciences
University of Pittsburgh

+Carroll School of Management
Boston College

Abstract **– Network connections are far from random and they have been shown to be correlated with external nodal attributes in a variety of cases. Therefore, metrics have been developed to quantify the extend of this phenomenon. In particular, the assortativity coefficient is used to capture the level of correlation between a single-dimensional nodal feature and the observed connections. However, in many cases, a vector representation of the node characteristics is more descriptive and provides a better understanding of the network structure. In this work, we develop a metric based on network randomization and empirical hypothesis testing that is able to quantify the assortativity patterns of a network with respect to a multi-dimensional node attribute. Our preliminary experimental results show that our metric outperforms a baseline extension of the assortativity coefficient, which has been previously used in the literature. We further showcase its applicability by using it to estimate the assortativity mixing of a social network dataset with respect to the mobility trails of its users.**

I. INTRODUCTION

Assortativity mixing refers to the phenomenon that describes the tendency of vertices to connect with each other when they present similar characteristics. These characteristics are either structural attributes (e.g., node degree) or external features (e.g., age of the node). Mixing patterns are important in complex network theory since they have a variety of implications. For instance, degree assortativity is closely related to the resilience of a network to targeted attacks [1]. Furthermore, mixing patterns with respect to external attributes have been integrated into generative network growth models [2,3].

The central idea behind quantifying assortativity patterns in a network is to compare the number of edges that connect nodes of *similar* type with the expected number of such connections if the latter were picked at random. This is exactly the basis of the assortativity coefficient [1] that can be applied for single dimensional attributes (numerical and categorical). In particular, if every node i is associated with a scalar value x_i, we can compute the normalized covariance of the values of x_i and x_j at the ends of an edge {i,j}. Then the assortativity coefficient is given by:

$$r = \frac{\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)x_i x_j}{\sum_{ij}\left(k_i\delta_{ij} - \frac{k_i k_j}{2m}\right)x_i x_j} \qquad (1)$$

where δ_{ij} is the Kronecker delta and k_i is the degree of node i. A similar expression exists for the case where node i is associated with a categorical attribute.

Nevertheless, formal treatment of mixing patterns for multidimensional nodal attributes has not received much attention [4], even though such scenarios appear in a variety of settings. For instance, in directed networks, the full degree of a node is represented by a two-dimensional vector, where each element represents either the in- or the out-degree of that node. Hence, if we do not want to lose any information during our analysis we need to consider the degree of a node as a vector [5]. Note also that vector attributes can describe complex behaviors, such as the purchase behavior of Amazon users, the urban mobility of city-dwellers etc.

Despite the prevalence of such settings, a formal metric for treating the assortativity with respect to a vector attribute is missing. The existing literature does not provide a metric that can be applicable in a generic scenario and is mainly focused on specific cases. For example, Foster *et al.* [5] defined 4 different types of degree assortativity for directed networks, essentially reducing the vector attribute to its elements. Block and Grund [6] utilized stochastic actor-oriented models for networks where the nodes have an increasing number of attributes in common. However, their approach is applicable to longitudinal and directed network data. In a tangential direction, Sanchez *et al.* [7] developed a method for statistical selection of congruent subspaces that have high dependency with the network structure. Recently, Pelechrinis [8] developed a generic metric for vector assortatitivity that projects the vectors to labels through a clustering process. However, given that clustering is known to be an ill-posed problem – at least under specific axiomatic frameworks [9,10] – selecting an appropriate clustering algorithm for all the cases might be impossible. Hence, its practical applicability is limited. Finally, the majority of the literature that deals with similar problems treats every element of the vector feature in isolation [11,12] – this will also serve as our baseline.

Contrary to the existing literature, we opt to provide a formal and generic metric for quantifying the assortativity patterns in a network with multi-attributed nodes. In a nutshell, our approach consists of a network randomization process, which allows us to estimate the pairwise similarity of connected nodes, if connections were made at random. Having this distribution allows us to perform an empirical hypothesis test, by comparing the actual similarity between connected nodes in the real network and the distribution of the randomization (null model). We evaluate our metric using synthetic data, where the ground truth is known. We further apply our metric on a dataset from a location-based social network and quantify the mixing patterns of this network with respect to the mobility traces of the users.

Roadmap: In Section II we introduce our proposed metric. Section III presents our preliminary evaluations, while Section V presents the application of our metric on a location-based social network. Finally, Section IV concludes our work and discusses its implications and our future directions .

II. OUR METRIC

To develop our metric we draw on the same intution that led to the development of the assortativity coefficient. In particular, we use a vector similarity metric ξ to compute the average similarity μ_ξ between connected nodes based on their vector attributes. Consequently, we bootstrap through network randomization the distribution for the average similarity $\mu_{\xi rand,}$ if connections were forming at random. This randomization can either be fully random (e.g., Erdos-Renyi random networks) or controled for specific network or external properties that are important to the setting at hand (e.g., configuration model). Once we have build the distribution for $\mu_{\xi rand}$ we can perform the following two-sided hypothesis test (at a predefined significance level α):

$$H_0: \quad \mu_{\xi rand} = \mu_\xi \quad (2)$$

$$H_1: \quad \mu_{\xi rand} \neq \mu_\xi \quad (3)$$

Failure to reject the null hypothesis essentially translates to random mixing in the network with respect to the vector attribute of the nodes. If the null is rejected, then the sign of the difference ($\mu_{\xi rand}$-μ_ξ) will inform us whether there is positive or negative mixing. However, the above test essentially responds to the question on whether the network is randomly mixed or not. We can further quantify the level of mixing. In particular, we compute the standardized mean difference d using the randomized network sample and transforming it to a value bounded between -1 and 1:

$$s = \frac{d}{\sqrt{d^2 + \varepsilon}} \quad (4)$$

where d is given by $d = \frac{\mu_\xi - \mu_{\xi rand}}{\sigma}$ and σ is the expected variance of the pairwise similarity in the randomized network. The latter can also be calculated through the repeated randomizations.

Equation (4) provides our final assortativity vector metric, with ε being a free parameter.

III. RESULTS

In order to evaluate our method we generate synthetic network data for which we know the ground truth with regards to the mixing patterns, in a manner similar to [8]. As alluded to above we compare with a baseline extension of the assortativity coefficient r (Equation (1)) that has been extensively used in the literature. In particular, we first calculate the assortativity coefficient r_i of the network with respect to each element i of the vector attribute. Then the baseline assortativity is given by:

$$r_{base} = \frac{\sum_{i=1}^{q} r_i}{q} \quad (5)$$

where q is the dimensionality of the nodal attribute. In our synthetic data we use q=5. Given that we know the ground truth for the mixing patterns in our networks our evaluation metric is the Root Mean Square Error (RMSE) of the assortativity values obtained from our metric and the baseline. Our metric further has two parameters that need to be chosen, namely, (a) the similarity metric ξ, and (b) ε in Equation (4). Therefore, we need to examine the sensitivity of our metric to these parameters. In particular, we consider three similarity metrics (cosine, correlation and Euclidean-based) and examine values of ε in [0.1,2].

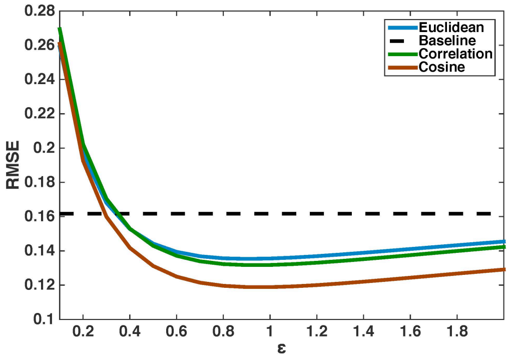

Figure 1 The proposed metric outperforms the baseline extension of assortativity coefficient. Furthermore, it does not appear sensitive to the choice of ε and/or similarity metric.

In Figure 1 we show that the performance our approach is very similar regardless of the specific similarity metric used. Moreover, the RMSE of our method is much lower compared to the baseline (besides very small values of ε).

IV. MOBILITY ASSORTATIVITY PATTERNS

Next we turn our attention to a real network dataset, and in particular, a dataset obtained from a location-based social network (LBSN), namely, Gowalla [13]. An LSBN consists of two components; (i) the social component that resembles any other digital social network, where users are connected based on "friendship" relations, and, (ii) the location component, which describes the mobility of the users based on their voluntary sharing of their whereabouts (through check-ins). Our dataset consists of 10,097,713 check-ins performed by 183,709 users in 1,470,727 distinct venues. Furthermore, there are 765,871 edges in the social (friendship) network.

Based on the above, every user u in this type of networks can be associated with a vector $\mathbf{c}_u$ that captures the places he has visited. In particular, the ith element of the vector is equal to the number of check-ins that u has in location/venue i. An important question that arises then is ``What are the assortativity patterns of this network with respect to the mobility trails of the users?''. The answer to this question has implications for the underlying spatial homophily of this network [14]. For answering this question we rely on our proposed metric (Section II), where we use the cosine similarity as our similarity metric. In particular, the similarity between users u and v is defined as:

$$\xi_{u,v} = \frac{\vec{c}_u \cdot \vec{c}_v}{\left\|\vec{c}_u\right\|_2 \left\|\vec{c}_v\right\|_2} \quad (6)$$

For our randomization we will consider two scenarios. First, we completely randomize the edges in the network, essentially sampling the G(n,m) Erdós-Rènyi random graph ensemble. Nevertheless, this will lead to an underestimation of the average pairwise similarity since the vast majority of (randomly selected) pairs will inevitably live in long distances and hence, the chances of having common venues visited will be small. Therefore, we will also perform a randomization where we will control for the distribution of the home-location distance of friends in the real network. Table 1 presents the computed average similarities for the real network as well as the 95% confidence interval from 100 instances of the two randomization processes. As we can notice the average pairwise similarity in the real network is significantly higher as compared to the one for the randomized networks. In particular, the average similarity in the real network is higher than the upper bound of the 95% confidence interval for both cases. It is also interesting to observe that the average similarity for the pure random graph network model is also significantly smaller as compared to the one in which we control for the home-location distance distribution of connected nodes.

Real network similarity	**ER network similarity**	**Home-location controlled random network similarity**
0.05425	[0.00233,0.0024]	[0.01834, 0.01837]

Table 1 There is a clear positive assortativity mixing with regards to the mobility trails of Gowalla users.

Using then equation (4) we can compute the coefficient, which is equal to **0.94** (p-value < 0.05), if we consider the pure ER network model as our baseline, and **0.31** (p-value < 0.05), if we control for the home-location distribution in our randomized baseline. As we can see the selection of the baseline model is really important and is application specific. For example, in the scenario examined it is clear (for the reasons also analyzed above) that the ER model overestimates the observed mixing patterns in the network.

V. CONCLUSIONS

In this work we designed an assortativity metric for multi-attributed networks. Our evaluations showed that our metric can quantify mixing patterns in the network more accurately than a baseline extension of the assortativity coefficient. We further showcased the applicability of our metric by computing the assorativity of location-based social network with respect to the mobility traces of the users.

We believe that our work can trigger more research on this largely under-represented topic and we hope to drive the development of related metrics for the emerging area of composite networks. The latter can be thought of as multidimensional networks with multiple types of edges and nodes. In such networks a direct application of metrics developed for unimodal networks will lead to large information loss [15]. For example, when there are multiple types of edges attached to a node, the degree of a node is not scalar anymore but a vector! Hence, using the assortativity coefficient to calculate the degree mixing of this network will ignore important information.

Acknowledgements: The work has been partially supported by the ARO YIP award #W911NF-15-1-0599 (67192-NS-YIP).

REFERENCES

[1] M.E.J. Newman (2003) Mixing patterns in networks. Phys. Rev. E 67, 026126.
[2] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri (2008) Feedback effects between similarity and social influence in online communities. In ACM SIGKDD
[3] M. Kin and J. Leskovec (2012) Latent multi-group membership graph model. In ICML.
[4] M.E.J. Newman (2010) Networks: An Introduction. Oxford University Press.
[5] J. Foster, D. Foster, P. Grassberger, and M. Paczuski (2010) Edge direction and the structure of networks. In Proceedings of the National Academy of Sciences, Vol. 107, No. 24 (May 2010).
[6] P. Block and T. Grund (2014) Multidimensional homophily in friendship networks. In Network Science, Vol. 2, No. 2 (Aug. 2014), pp. 189-212.
[7] P.I. Sanchez, E. Muller, F. Laforet and F. Keller (2013) Statistical Selection of Congruent Subspaces for Mining Attributed Graphs. In IEEE ICDM.
[8] K. Pelechrinis (2014) Matching patterns in networks with multi-dimensional attributes: a machine learning approach. In Social Network Analysis and Mining, Vol. 4, No. 1 (Apr. 2014), pp. 1-11.
[9] L. Fisher and J. W. V. Ness (1971) Admissible clustering procedures. In Biometrika, vol. 58, no. 1, pp. 91-104.
[10] J. Kleinberg (2002) An impossibility theorem for clustering. In NIPS.
[11] H. Lauw, J. Shafer, R. Agrawal, and A. Ntoulas (2010) Homophily in the digital world: A LiveJournal case study. In IEEE Internet Computing, Vol. 14, No. 2, pp. 15-23.
[12] K. Zhao, L. Ngamassi, J. Yen, C. Maitland, and A. Tapia (2010) Assortativity patterns in multi-dimensional inter-organizational networks: a case study of the humanitarian relief sector. In SBP.
[13] E. Cho, S.A. Myers and J. Leskovec (2011) Friendship and Mobility: User Movement in Location-based Social Networks. In ACM SIGKDD
[14] K. Zhang and K. Pelechrinis (2014) Understanding Spatial Homophily: The Case of Peer Influence and Social Selection. In ACM WWW.
[15] Y. Sun and J. Han (2013) Mining heterogeneous information networks: a structural analysis approach. ACM SIGKDD Explorations Vol. 14, No. 2, pp. 20-28.