Structural diversity based on network embeddings

Styliani Bourli [1], Kouzouglidis Panagiotis [2]

Department of Computer Science and Engineering, University of Ioannina, Greece
E-mail: [1] sbourli@cs.uoi.gr, [2] pkouzouglidis@cs.uoi.gr

Abstract

Network embeddings have gained huge popularity in the recent years as a powerful tool to analyze social networks. A commonly used metric in social network analysis is diversity. Diversity of a node expresses how un-similar are the neighbors in the network. In this paper we try to measure node diversity in a network using network embeddings. For this purpose we suggest an embedded-based definition which measures node diversity. We perform a number of extensive experimental measurements in 5 real-world and 2 synthetic networks. We study the correlation between node diversity produced by our definition using embeddings from different embedding methods. Moreover we search if exist any correlation between diversities produced by our definition and other known network metrics. We find that the most of the embedding methods are correlated between them, as well as that diversities of our definition are correlated with some of the network metrics.

Keywords: embeddings, diversity, metrics, networks

1. Intoduction

Embeddings express the transformation of nodes, edges, and features of a network into a vector space of low-dimension. These representations can be used for a wide range of tasks on networks such as classification, clustering, link prediction, and visualization. They make it easier to do machine learning on large inputs and also vector operations are simpler and faster than comparable operations on graphs. This make's them a useful tool for social network analysis.

In this paper we give an embbedding-based definition for a known metric of social network analysis, which called diversity and measure how un-similar is a node compared to its neighbors. Generally, diversity in real networks tend to be low. Diversity between linked nodes in a network is low, when most of the nodes are connected with similar ones. This happens because people tend to create links with those, who have similar interests, opinion or beliefs with them. Low diversity leads people to live in a "world" dominated by their own preferences and the preferences of their neighbors. This makes them lose the feeling of what is generally in the "world". For this reason diversity is a very interesting metric.

Embeddings used in our definition can be produced with many different embedding methods. We select four of them and we study the correlation between the diversities produced by our definition using the four different embedding representations. Another correlation we study is the correlation between the diversities of our definition, using embeddings from the different embedding methods, and other metrics, that used in social network analysis, such degree, PageRank, clustering coefficient and structural diversity.

For the purpose of our experiments we use 5 real and 2 synthetic networks. We find out that in most of the cases, three of the four methods have high correlation in diversities produced using their embeddings. We see also that diversity is correlated with some of the other metrics examined.

2. Embedding methods

Embeddings integrate information about network in a lower dimension vector space. This makes them very important in social network analysis. We can learn embeddings across network embedding methods. There is a wide variety of embedding methods. In this paper we focus on four of them, Node2Vec, DeepWalk, VERSE and GraRep. They train a neural network using random walks -created with different techniques- in order to produce embeddings.

- **Node2vec.** Node2Vec [1] learns node embeddings- d-dimensional representations -by training a neural network with biased random walks. There are two parameters p, q that control if the random walks follow BFS or DFS strategy. Parameter p controls the likelihood of immediately revisiting a node in the walk. Setting it to a high value (> max(q, 1)) ensures that we are less likely to sample an already visited node in the following two steps of the random walk. On the other hand, if p is low (< min(q, 1)), it would lead the walk to backtrack a step and this would keep the walk "local". Parameter q allows the search to differentiate between "inward" and "outward" nodes. if q > 1, the random walk is biased towards nodes

close to node t. Such walks obtain a local view of the underlying graph with respect to the start node in the walk and approximate BFS behavior in the sense that our samples comprise of nodes within a small locality. In contrast, if q < 1, the walk is more inclined to visit nodes which are further away from the node t. Such behavior is reflective of DFS which encourages outward exploration.

- **DeepWalk.** DeepWalk [2] learns node embeddings- d-dimensional representations -by training a neural network with uniform random walks starting from each node. The sampling strategy in DeepWalk can be seen as a special case of node2vec with p=1 and q=1.

- **VESRE.** VERtex Similarity Embeddings (VERSE) [3] is a simple, versatile, and memory-efficient method that derives graph embeddings explicitly calibrated to preserve the distributions of a selected vertex-to-vertex similarity measure.

- **GraRep.** GraRep [4] learns low dimensional vectors to represent vertices appearing in a graph and, unlike the aforementioned methods, integrates global structural information of the graph into the learning process. More specifically, it suggests using Singular Value Decomposition (SVD) on a log-transformed DeepWalk transition probability matrix of different orders, and then concatenate the resulting representations.

3. Diversity measurement based on embeddings

In this section we propose a measurement, that uses embedding representations of the nodes to compute their diversities. We define the diversity of a node to be equal to the average distance between the embeddings of its neighbors. Since embeddings are expressed as vectors, we choose *cosine similarity* as a similarity metric between two node embeddings. Let two nodes of the network *u, v* and their corresponding embedding representations e_u, e_v. Cosine similarity of e_u, e_v takes values in range [0,1] and is defined as:

$$similarity(e_u, e_v) = \frac{e_u \cdot e_v}{\|e_u\| \|e_v\|} \quad (1)$$

As a result the distance between e_u, e_v using (1) is:

$$distance(e_u, e_v) = 1 - similarity(e_u, e_v) \quad (2)$$

Eventually, the diversity of a single node *u* from (1),(2) is defined as:

$$diversity(u) = \frac{1}{|N(u)|} \sum_{v \in N(u)} distance(e_u, e_v) \quad (3)$$

where N(u) expresses the neighborhood of u.

4. Network metrics

There is a wide number of metrics that are used in social network analysis. In this paper we care about metrics that are related to the nodes of the network. Some of them are degree, PageRank and clustering coefficient. More specifically, let G = (V, E) be a given network, where V are the nodes of the network, and E are the edges, E ⊆ (V × V):

- **Degree** of a node. Degree of a node is the size of node's neighborhood. Let uϵV and N(u) is the set of neighbors of u, then d(u) = |N(u)|

- **PageRank.** PageRank is a link analysis algorithm which assigns a numerical weight to each node of the network. This weight expresses how important is the node. Let u,vϵV, d(u) is the degree of node u, aϵ(0,1) is the transition probability - $\alpha = 0.85$ in most cases - and w_u is the weight of node u. Then,

$$W_v = a \sum_{u \to v} \frac{1}{d(u)} w_u + (1-a)\frac{1}{n}$$

  where n is the number of network nodes

- **Clustering coefficient.** Clustering coefficient (C) is a measure of the degree to which nodes in a graph tend to cluster together:

$$C = \frac{number\ of\ closed\ triplets}{number\ of\ all\ triplets\ (open \wedge closed)}$$

An also known metric in social network analysis is **structural diversity** as it is defined in paper [5]. Structural diversity of a node expresses how different is a node compared to its neighbors by a defined measure. More specifically, in [5] structural diversity of a node is measured as the number of connected components in the subgraph which consists of the neighbors of the node. For large neighborhoods, only connected components, that have size bigger than a certain number k are taken into account.

5. Experiments

In this section, we first describe the datasets that we choose and how we use them. Then we present the qualitative and the quantitative results of our experiments.

For our experiments we use five real and two synthetic networks, which we analyze below. For each of the networks, that isn't connected, we select and use for our measurements only the largest connected component. Then we compute node embeddings for each graph using different embedding methods and use them to measure diversity with our definition. We also compute the network metrics described in *section 4*. For Node2Vec method we use parameters p=2 and q=1/2, because we don't want random walks to revisiting a node and also we want to follow DFS behavior. All the other parameters of the methods set as the default, except the dimension of the vectors, which we set size 64. For structural diversity metric we take into account only connected components, which have size bigger than 2.

In the end of the section are presented the results and the correlation between them.

5.1 Datasets

For the purposes of our experiments we choose five real networks from SNAP and we also create two synthetic networks.

NetWork	#nodes	#edges	avg degree
Ego-facebook	4039	88234	43.6910
Email-Eu-core	986	16687	33.8479
Ca-GrQc	4158	13428	6.4589
P2p-Gnutella06	8717	31525	7.2330
Wiki-Vote	7066	100736	28.5129
Preferential attachment	1005	15824	31.4905
Erdős–Rényi	1005	17063	33.9562

Table 1: Information of the largest connected component of each network.

Table 1 shows some statistical information for each of the networks. Some interesting information about them exist below.

Ego-facebook: an ego-network, where nodes are Facebook users and edges are friendships between them
Email-Eu-core: a network, where nodes are members of a large European research institution and an edge between two members exist if they have send each other at least one email
Ca-GrQc: a network, where nodes are authors submitted to General Relativity and Quantum Cosmology category and an edge exist between two authors if they have co-authored a paper
P2p-Gnutella06: a network, where nodes represent hosts in the Gnutella network topology and edges represent connections between them
Wiki-Vote: a network, where nodes are users participating in elections for voting a Wikipedia administrator and edges are votes between users
Preferential attachment: a network created by barabasi_albert_graph method from NetworkX library given 1005 as first parameter (number of nodes) and m = 17000/1005 as second parameter (number of neighbors for each new node)
Erdős–Rényi: a network created by erdos_reyni_graph method of NetworkX library given 1005 as first parameter (number of nodes) and 0.034 as second parameter (probability of edge creation)

5.2 Results

In this subsection we compare the diversity results of the diversities produced by the different embedding methods from *section 2*. Also, we present the correlation of our diversity measurement with the other network metrics described in *section 4*.

5.2.1 Diversity correlation of the embedding methods

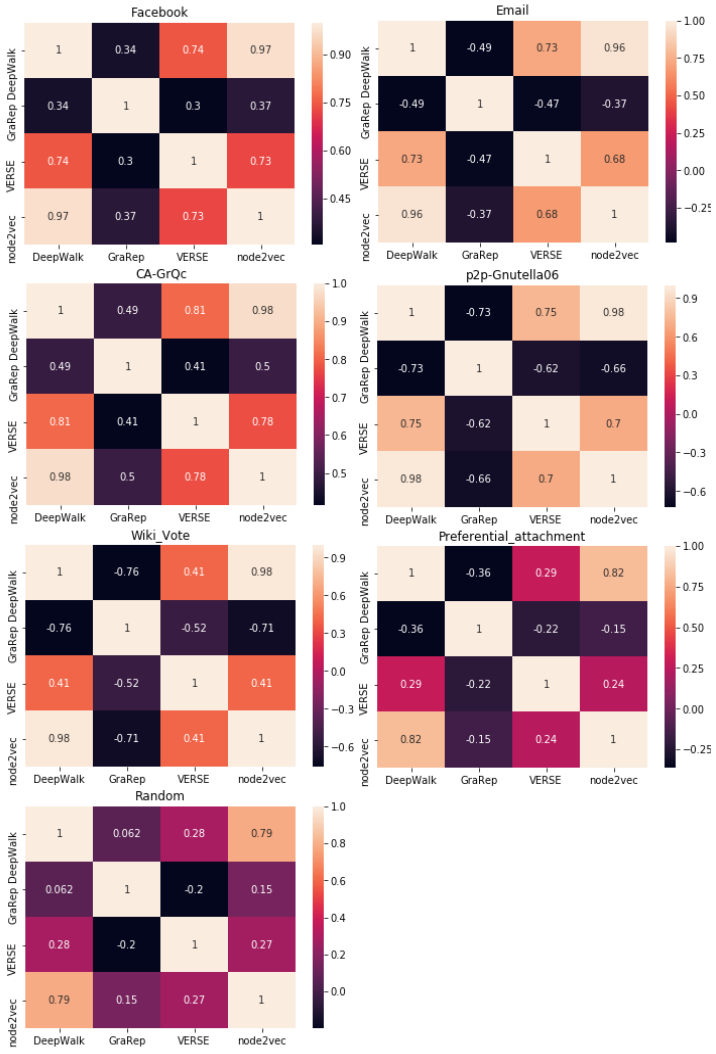

Figure 1: Correlation between diversities produced by different embedding methods for each of the networks.

Here we study the correlation between diversities produced with embeddings from the different embedding methods. To express the linear correlation between the diversities, we use *Pearson correlation coefficient (PCC)* measure. It takes values in the range [-1, 1], where value 1 is total positive linear correlation, 0 is no linear correlation and -1 is total negative linear correlation.

In *Figure 1*, are presented the correlations between diversities for the different embedding methods for our seven networks. For the synthetic networks, we created 10 random networks for each case, we computed the correlation between the embedded based diversities and we took the average.

We observe that Node2Vec and DeepWalk methods produce diversities that have high correlation in the most of the cases. This is reasonable since the methods produce embeddings with a quite similar way.

It is also obvious that diversities from VERSE embedding method are correlated with them from DeepWalk and Node2vec methods. This means that the produced node embeddings of this method are similar with the node embeddings from the other methods.

Unlike the rest of the methods, diversities from GraRep method seem to have very low correlation values compared with the other results. This means that the produced embeddings from this method have significant difference in relationship with the others.

Finally, in all of the cases the correlation values that refer to the real networks are higher than the values refer to the synthetic networks.

DeepWalk	Structural diversity	Degree	PageRank	Clustering coefficient
Ego-facebook	-0.3011	**0.0436**	0.0049	-0.4824
Email-Eu-core	0.2560	**0.5331**	0.5132	-0.2037
Ca-GrQc	0.1966	0.2063	**0.2909**	-0.3602
P2p-Gnutella06	0.3762	**0.7501**	0.7140	0.0314
Wiki-Vote	**0.5956**	0.3981	0.3862	0.1750
Preferential attachment	-0.0209	**0.0532**	0.0530	-0.0091
Erdős–Rényi	0.0192	**0.0436**	0.0432	-0.0465

Table 2: Correlation between diversities produced by embeddings from DeepWalk method and other network metrics.

Node2vec	Structural diversity	Degree	PageRank	Clustering coefficient
Ego-facebook	-0.2998	-0.0632	**-0.0299**	-0.4815
Email-Eu-core	0.1168	**0.4002**	0.3811	-0.2288
Ca-GrQc	0.1369	0.1659	**0.2375**	-0.3855
P2p-Gnutella06	0.3731	**0.7248**	0.6808	0.0358
Wiki-Vote	**0.5308**	0.3448	0.3316	0.1474
Preferential attachment	-0.0132	**0.0325**	0.0323	-0.0189
Erdős–Rényi	0.0229	**0.0291**	0.0289	-0.0566

Table 3: Correlation between diversities produced by embeddings from Node2Vec method and other network metrics.

VERSE	Structural diversity	Degree	PageRank	Clustering coefficient
Ego-facebook	-0.3333	0.0860	**0.1161**	-0.5403
Email-Eu-core	0.2589	**0.4903**	0.4818	-0.0899
Ca-GrQc	0.3388	0.2335	**0.3837**	-0.2893
P2p-Gnutella06	0.2651	**0.5762**	0.5612	0.0334
Wiki-Vote	**0.3213**	0.2582	0.2660	0.1308
Preferential attachment	-0.0106	0.0356	**0.0358**	-0.0053
Erdős–Rényi	0.0118	**0.0452**	0.0451	-0.0160

Table 4: Correlation between diversities produced by embeddings from VERSE method and other network metrics.

GraRep	Structural diversity	Degree	PageRank	Clustering coefficient
Ego-facebook	-0.3316	-0.1808	**-0.0005**	-0.1097
Email-Eu-core	-0.6513	-0.3479	-0.3308	**-0.2591**
Ca-GrQc	**0.0259**	-0.1893	0.0165	-0.5064
P2p-Gnutella06	-0.3822	-0.6181	-0.5908	**-0.1028**
Wiki-Vote	-0.7700	-0.5131	-0.5069	**-0.3566**
Preferential attachment	**0.0298**	-0.0514	-0.0504	-0.0541
Erdős–Rényi	**0.0067**	-0.0559	-0.0547	-0.0377

Table 5: Correlation between diversities produced by embeddings from GraRep method and other network metrics.

5.2.2 Correlation between embedding-based diversity and other metrics

Here we study the correlation of diversities produced via our diversity embedding definition with other network metrics for each of the seven networks. Especially, we produce a matrix with the correlations between the metrics and our diversity for each of the four different embedding methods.

Tables 2, 3 and *4* refer to diversity correlations produced with embeddings from DeepWalk, Node2Vec and VERSE, respectively. We observe that our embedding definition has great correlation with degree and PageRank metrics. More specifically, the degree method has the greater correlation values in most of the networks. Structural diversity metric doesn't show similar behavior with the other metrics except the case of Wiki-Vote network, where the correlation with

our diversities is high. Finally, clustering coefficient has negative correlation values in most of the cases, something shows that has not particular correlation with diversities of our definition.

On the other hand, GraRep differs enough from the other metrics. More specifically, it presents a small correlation with structural diversity metric and negative correlation values with all of the other metrics. It is expected, because GraRep, as we saw in *subsection 5.2.1*, hasn't any correlation with the other three methods.

An important conclusion we can extract from the above is that the degree and the PageRank metrics have correlation with our embedding diversity definition in most of the cases. This means that the size of the neighborhood and the importance of a node, is analogous to its diversity. So if a node has a great number of neighbors, the node is quite different from them. Respectively, as greater PageRank a node has, as different this node is from its neighbors. So if we find the top nodes having the greatest PageRank or the greatest degree, then it very likely that these nodes have high diversity.

6. Conclusions and Future Work

In this paper, we presented a definition that computes the diversity of network nodes based on their embedding representations. We used for this purpose four well known embedding methods and studied the correlation between the diversity that produced by our definition using the embeddings of the methods. We showed that Node2Vec, DeepWalk and VERSE produce diversity values that are highly correlated, unlike with GraRep, which didn't have correlation with any of the above methods. For this reason we concluded that the embedding representations of Node2Vec, DeepWalk and VERSE are quite similar. Another correlation we measured was this between diversities of our definition and other network metrics; degree, PageRank, clustering coefficient and structural diversity from [5]. From our measurements we came to the conclusion that diversity in many real networks has correlation with degree and PageRank metrics. This made important nodes or nodes with great neighborhood, also nodes with high diversity.

A future work could be to give another definition of diversity using different similarity measures except cosine similarity and measure the correlation between the different embedding methods. We could also compute the correlation between the given definition of diversity and other network metrics except the examined. Finally, we could see if our conclusions apply also to other graphs than the ones we chose.

7. References

[1] Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks, KDD '16, August 13 - 17, 2016, San Francisco. CA, USA

[2] Bryan Perozzi, Rami Al-Rfou and Steven Skiena. DeepWalk: Online Learning of Social Representation. ACM 2014

[3] Anton Tsitsulin, Davide Mottin, Panagiotis Karras and Emmanuel Müller. VERSE: Versatile Graph Embeddings from Similarity Measures. s. In WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France. ACM, New York, NY, USA

[4] Shaosheng Cao, Wei Lu and Qiongkai Xu. GraRep: Learning Graph Representations withGlobal Structural Information. CIKM'15, October 19–23, 2015, Melbourne, Australia

[5] Xin Huang, Hong Cheng, Rong-Hua Li, Lu Qin, Jeffrey Xu Yu: Top-K structural diversity search in large networks. VLDB J. 24(3): 319-343 (2015)