Network Similarity via Multiple Social Theories

Michele Berlingerio
IBM Research Dublin
Email: mberling@ie.ibm.com

Danai Koutra
Carnegie Mellon University
Email: danai@cs.cmu.edu

Tina Eliassi-Rad
Rutgers University
Email: eliassi@cs.rutgers.edu

Christos Faloutsos
Carnegie Mellon University
Email: christos@cs.cmu.edu

***Abstract*—Given a set of k networks, possibly with different sizes and no overlaps in nodes or links, how can we quickly assess similarity between them? Analogously, are there a set of social theories which, when represented by a small number of descriptive, numerical features, effectively serve as a "signature" for the network? Having such signatures will enable a wealth of graph mining and social network analysis tasks, including clustering, outlier detection, visualization, etc. We propose a novel, effective, and scalable method, called NETSIMILE, for solving the above problem. Our approach has the following desirable properties: (a) It is supported by a set of social theories. (b) It gives similarity scores that are size-invariant. (c) It is scalable, being linear on the number of links for graph signature extraction. In extensive experiments on numerous synthetic and real networks from disparate domains, NETSIMILE outperforms baseline competitors. We also demonstrate how our approach enables several mining tasks such as clustering, visualization, discontinuity detection, network transfer learning, and re-identification across networks.**

I. INTRODUCTION & PROPOSED METHOD

We address the problem of *network similarity*. Specifically, given a set of k networks of potentially different sizes and without any assumptions on overlapping nodes or edges, how can we efficiently provide a meaningful measure of structural similarity (or distance)? For example, how structurally similar are the ASONAM and ICWSM co-authorship networks? How does their structural similarity compare with the similarity between the ASONAM and WSDM co-authorship networks? Such measures are extremely useful for numerous social network analysis and graph mining tasks. One such task is clustering: given a set of networks, find groups of similar ones; conversely, find anomalies or discontinuities – i.e., networks that stand out from the rest. Transfer learning is another application. If networks G_1 and G_2 are similar, we can transfer conclusions from one to the other to perform across-network classification with better predictive accuracy.

When considering the problem of network structural similarity, we need to make some choices. Should the comparison be at the local (node) level, at the global (network) level, or both? Should the comparison be based on the similarities (or distances) of the adjacency matrices or similarities (or distances) of structural features, or both? Should the approach be interpretable or is a black-box approach okay? Must the approach be scalable? Can the approach be readily extended to accommodate non-structural features? Clearly, these choices are not independent of each other. For example, comparisons at the local level tend to be more interpretable and scalable.

We present an approach, NETSIMILE, that has the following seven characteristics. (1) It can compare networks at the local (node and neighborhood) level. (2) It uses structural features supported by social theories. (3) It is scalable. (4) It is interpretable. (5) It is size-independent. (6) It can readily be extended to accommodate global-level features and non-structural features. (7) Its similarity values satisfy the *Identity*, *Symmetry*, and *Divergence* properties.

The core of NETSIMILE is a careful extraction, aggregation, and evaluation of structural features from nodes and their local neighborhoods. For every network G, NETSIMILE derives a small number of numerical features, which incorporate various social theories and capture the topology of the network as moments of distributions over its local structural features. Specifically, NETSIMILE extracts the following features for every node: *degree*, *clustering coefficient*, *average degree of neighbors*, *average clustering coefficient of neighbors*, *number of edges in ego-network*, *number of outgoing edges of ego-network*, and *number of neighbors of ego-network*. NETSIMILE then applies these aggregator functions on each local feature to generate the "signature" vector for a graph: *median*, *mean*, *standard deviation*, *skewness*, and *kurtosis*. The similarity score between two networks then is just the similarity of their "signature" vectors. Once we have the similarity function, we can do a wealth of interesting tasks, including clustering, visualization, anomaly detection, etc.

NETSIMILE incorporates four social theories when extracting the "signature" vector of a network: *Social Capital*, *Structural Hole*, *Balance*, and *Social Exchange*. These theories, respectively, capture connectivity of nodes and their neighborhoods, control of information flow, transitivity among the nodes, and reciprocity among the nodes. We selected these social theories because they are purely structural (as supposed say homophily which relies on a non-structural characteristic). Also, these theories can be applied to a wide range of networks as opposed to just social networks.

Our empirical study includes experiments on more than 30 real-world networks and various synthetic networks generated by four different graph generators (namely, Erdös-Rényi, Forest Fire, Watts-Strogatz, and Barabási Preferential Attachment). We compare NETSIMILE with two baselines. The first baseline, FSM, extracts frequent subgraphs from the given graphs and performs pairwise comparison on the intersection of the two sets of frequent patterns. The second baseline, EIG, computes the *k* largest eigenvalues of each network's adjacency matrix and measures the distance between them.

Our experiments provide answers to the following questions: How do the various methods compare w.r.t. their similarity scores? Are their results intuitive (e.g., is a social network more similar to another social network than to a technological network)? How do they compare to null models? Are the methods just measuring the sizes of the networks in their

comparisons? How scalable are the various methods?

Due to brevity, we have omitted a full description of NETSIMILE and all of our experiments here. The reader can find these in [1].

II. EXPERIMENTS

Figure 1 shows the NETSIMILE similarity scores between various networks (described in [1]) and their rewired versions. NETSIMILE similarity score (i.e. 1 minus the NETSIMILE Scaled Canberra Distance) is in $[0, 1]$, with 0 meaning no similarity at all and 1 meaning identical graphs. We rewire a graph by randomly reassigning a number of its edges. The rewiring parameter, $c \in [0, 1]$, determines the fraction of edges rewired. Edges are reassigned in a way that preserves the expected degree of each node in the graph. When c is 0, no rewiring takes place (i.e., the original and rewired graphs are identical) and as expected the NETSIMILE score is 1 between them. When c is 1, the rewired graph is the least similar to the original graph (because all the edges in the original graph have been randomly reassigned). As expected, comparing an Erdös-Rényi graph to its rewired version does not significantly change the NETSIMILE score (see the line with the black circles in Figure 1). However, for real-world graphs (like co-authorship networks and autonomous systems networks) as the rewiring parameter increases, the NETSIMILE score decreases.

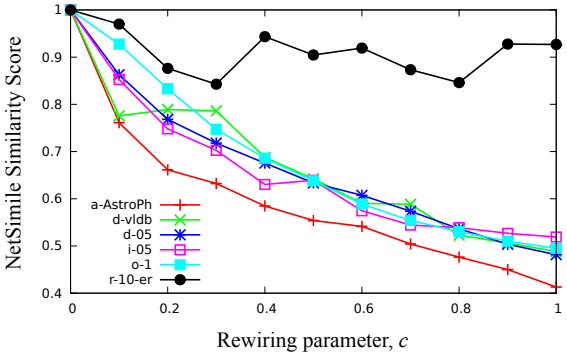

Fig. 1. NETSIMILE similarity scores for various graphs and their rewired versions. As the rewiring parameter increases, the NETSIMILE similarity score decreases in real-world networks (e.g., co-authorship networks from arXiv, DBLP, and IMDB and technological networks from Oregon AS). Unsurprisingly, increasing the rewiring parameter in an Erdös-Rényi graph (black circles) does not have the same pattern.

NETSIMILE as a Measure of Node-Overlap. Given three graphs G_A, G_B, and G_C of the same domain (e.g., co-authorship networks in *SIGMOD*, *VLDB* and *ICDE*), can we use *only* their NETSIMILE's "signature" vectors to gauge the amount of node-overlap between them? Our hypothesis is that if graph G_A is more similar to graph G_B than graph G_C, then G_A will have more overlap in terms of nodes with G_B than G_C. To test this hypothesis, we ran NETSIMILE with Canberra Distance on our real networks. Figure 2(a) depicts the scatterplot of NETSIMILE results on graphs within each comparable group (i.e., arXiv, DBLP-C, DBLP-Y, IMDb, Query Log, and Oregon AS graphs). The y-axis is the normalized node overlap and is equal to $\frac{|V_{G_A} \cap V_{G_B}|}{\sqrt{|V_{G_A}| \times |V_{G_B}|}}$. As the figure shows the lower the NETSIMILE Canberra Distance, the higher the normalized node intersection. This confirms our hypothesis that NETSIMILE can be used to gauge node-overlap between two graphs without node correspondence information. Figure 2(b) shows the same scatter plot, but computed using the EIG Canberra Distance approach. In this case, there is no correlation between node overlap and the distance. Due to its scalability issues, the FSM approach could not be computed on all the networks in Figure 2.

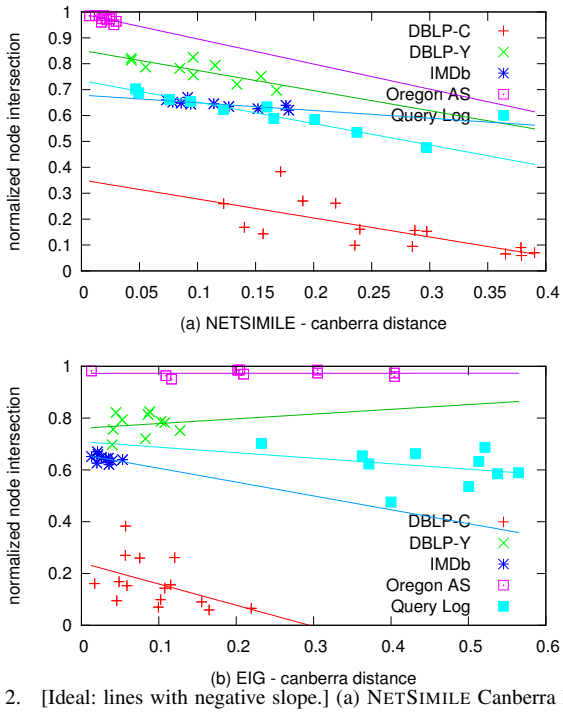

Fig. 2. [Ideal: lines with negative slope.] (a) NETSIMILE Canberra Distance on DBLP, IMDb, Oregon and QueryLog. (b) EIG Canberra Distance on the same networks. NETSIMILE is an effective measure for node overlap without any node-correspondence information. The lower the NETSIMILE Canberra Distance, the higher the normalized node intersection. This correlation does not hold for EIG. The points in both plots are along the fitted lines. For NETSIMILE (a), the root mean square (RMS) of residuals are $6.5E{-}2$ for DBLP-C, $2.6E{-}2$ for DBLP-Y, $9.0E{-}3$ for IMDb, $1.4E{-}2$ for Oregon AS, and $6.5E{-}2$ for Query Log. For EIG (b), the RMS of residuals are $8.2E{-}2$ for DBLP-C, $4.2E{-}2$ for DBLP-Y, $1.3E{-}3$ for IMDb, $1.2E{-}2$ for Oregon AS, and $6.7E{-}2$ for Query Log.

III. CONTRIBUTIONS

The contributions of our work are as follows. (1) *Novelty*: By using moments of distribution as aggregators, NETSIMILE generates a single "signature" vector for each graph based on the local and neighborhood features of its nodes. Our features incorporate four social theories that are endogenous to the network and are applicable to more than just social networks. (2) *Effectiveness*: NETSIMILE produces similarity / distance measures that are size-independent, intuitive, and interpretable. The similarity values satisfy the identity, symmetry, and divergence properties. (3) *Scalability*: The runtime complexity for generating NETSIMILE's "signature" vectors is linear on the number of edges. (4) *Applicability:* NETSIMILE's "signature" vectors are useful in many social network analysis and graph mining tasks.

REFERENCES

[1] M. Berlingerio, D. Koutra, T. Eliassi-Rad, and C. Faloutsos, "Netsimile: A scalable approach to size-independent network similarity," *CoRR*, vol. abs/1209.2684, 2012. [Online]. Available: http://arxiv.org/abs/1209.2684