NETWORK DISCOVERY USING CONTENT AND HOMOPHILY

*Steven T. Smith, Rajmonda S. Caceres, Kenneth D. Senne, Molly McMahon, Timothy Greer**

MIT Lincoln Laboratory; 244 Wood Street; Lexington MA USA 02420
{ stsmith, rajmonda.caceres, senne, molly.mcmahon, timothy.greer }@ll.mit.edu

ABSTRACT

A new approach for targeted graph sampling is proposed in which graph sampling and classification occur together, and content-based homophily is exploited to achieve improved classification performance. The application of network discovery of relevant content is considered using an approach that may be generalized to a broad class of vertex properties. The resulting procedure provides the initial step of a graph analytic processing chain whose performance is directly affected by the quality of graph sampling. The performance of the algorithm is measured with real network data and content observed on a social media site. Precision-Recall performance improvements of 30% are demonstrated with this dataset, compared to a baseline approach that does not exploit homophily. Because real-world graphs grow exponentially, this performance improvement may have a significant impact on graph analytic algorithms with sensitivities to the graph sampling quality.

1. INTRODUCTION

The challenging steps in the processing chain for graph analytics are the collection of relevant data, detection of important subgraphs, and the inference of vertex properties. The details of the initial graph sampling design have important consequences for all subsequent tasks; however, this step is oftentimes approached with ad hoc methods with detrimental effects. Furthermore, and especially in the context of flourishing social media and online content, effective graph sampling methods are essential to limit the quantity of irrelevant data and focus analysis.

Figure 1 shows the growth of neighborhood size as a function of number of hops when the graph is explored using the breadth-first search procedure. As shown in this figure, exploring social media graphs without prioritization leads to exponential growth of neighborhood size. This phenomena is in stark contrast with our understanding that true social communities have relatively small constant sizes, about 150 or so based on the well-established Dunbar result [15]. These communities become detectable only when the amount of irrelevant data is reduced to a tolerable size. Therefore, it is critical that data sampling techniques reduce the amount of noise introduced by indiscriminate sampling.

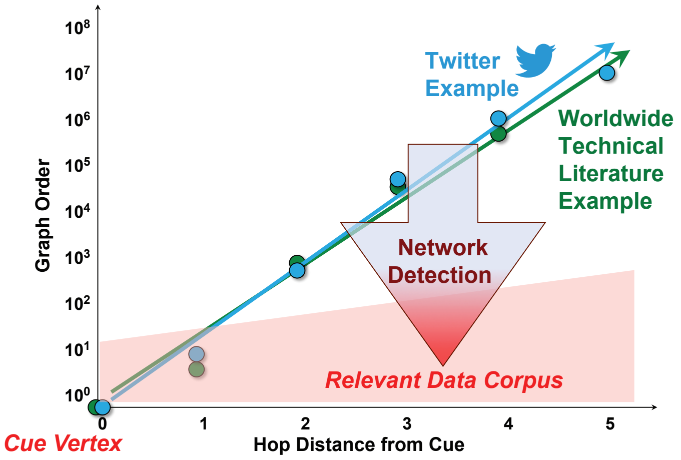

Fig. 1. Graph order versus hop length on real data using representative examples from a social media network and technical publications. The graph size grows exponentially; however, a small fraction of graph vertices contain relevant content.

The approach presented here is focused precisely on addressing sampling issues. Targeted graph sampling is proposed in which graph sampling and classification occur together, and content-based homophily is exploited to achieve improved classification performance. Notions of relevance are incorporated via a trained classifier, which then supervises and prioritizes sampling at each step. The performance of the algorithm is measured with real network data and content observed on a social media site. The resulting procedure provides the initial step of a graph analytic processing chain that includes the detection of important subgraphs and inference of vertex properties by using algorithms whose performance is directly affected by the quality of graph sampling.

Traditional graph sampling techniques [1, 3, 4, 6] focus on the goal of generating a subgraph, usually unbiased, whose topological characteristics preserve those of the original, larger graph. Depending on the learning task, different topological properties and different portions of the graph might be relevant. However, for the problem of discovering a specific network that is relevant to a particular subject, graph sampling not designed to include these specifics fails to discover the objective net-

*This work is sponsored by the Assistant Secretary of Defense for Research & Engineering under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

works [2]. Our approach leverages a trained classifier to bias the sampling procedure toward only the most relevant portions of the graph based on a priori domain knowledge of relevance.

The central assumption of our sampling approach is homophily between vertices, specifically common content of interest to relevant vertices. An existing approach that exploits content-based homophily is provided by Şimşek and Jensen [11] who consider the task of finding shortest paths to a target node and utilize both homophily and degree centrality to bias the sampling towards the nodes on the shortest path. We also leverage homophily to drive the sampling process; however, our learning objective is focused on relevance-based prioritization of vertices rather than distance-based prioritization.

Several authors have looked at semi-supervised vertex classification [2, 5, 12, 14], where the label information on a few nodes and the topology of the graph is used to infer the unknown labels of the rest of the nodes. Smith et al. [12] and Fang et al. [2] consider the important real-world setting in which graph sampling and classification occur together and locally. This is precisely the problem setting we consider. Our contribution is to combine the assumption of content-based homophily with this approach to achieve improved performance for both local graph sampling and classification. Practically, this is straightforward to accomplish because local neighborhood content is easily incorporated into the classifier, a design motivated by the assumption that shared affiliation is reflected in similar content among neighbors—homophily.

2. MATHEMATICAL MOTIVATION

The ultimate objective of a targeted graph sampling approach is the discovery of relevant network activity. Though homophily is a well-established phenomenon, this property has not been widely utilized for targeted graph sampling and cued network detection. In this section, we extend an existing formulation for optimum network detection to incorporate content-based homophily. Neyman-Pearson optimal network detection of a subgraph within a graph is achieved by thresholding the association probabilities of each vertex in the graph with a set of cue vertices [12]. This algorithm maximizes the probability of detecting the subgraph of vertices for a fixed false alarm probability, and is a consequence of diffusion via random walks on the graph. The novel contribution of this article is to exploit homophily for graph diffusion and hence network affiliation and detection.

An optimum network detection approach that employs homophily has the following straightforward representation. Let $G = (V, E)$ be an irreducible graph (i.e. G is strongly connected) with probabilities of relevance $\theta_1, \ldots, \theta_C$ given at the observed vertices $v_1, \ldots, v_C$. Let ψ_v be the probability that relevance propagates through vertex v to its neighbors. Otherwise relevance propagates to an absorbing "non-relevant" state with probability $1 - \psi_v$. Bayes' rule implies that the relevance at graph vertices is determined by a random walk with Markov transition matrix t_{vu} from vertex v to u, and the a priori probability *relevance diffusion model* ψ_v that represents that relevance propagates through v.

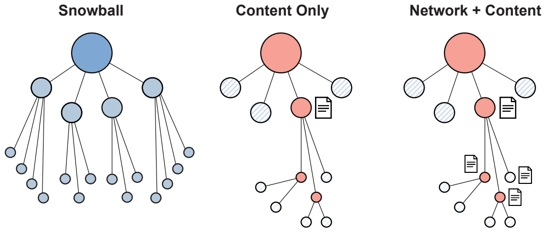

Fig. 2. Targeted graph sampling. The baseline sampling approach on the left explores every vertex encountered. The sampling in the center uses the vertex's content to determine relevance. Our approach on the right uses both vertex and neighborhood content to determine relevance.

Homophily is modeled by setting the transition probabilities to be proportional to a measure of content similarity between vertices:

$$t_{vu} \propto \text{similarity}(\text{content}_u, \text{content}_v). \tag{1}$$

An algorithm that quantifies similarity between vertices will be detailed in Section 3. Various relevance diffusion models are specified by path length from cue vertices and time-dependent kernels [12]. Probabilities of relevance at each vertex are determined by the *relevance propagation equation:*

$$\theta_v = \psi_v \sum\nolimits_{u \in N(v)} t_{vu} \theta_u, \tag{2}$$

which is the average of the neighboring relevance probabilities weighted by transition probabilities [12].

3. ALGORITHM

Start with a small set of example vertices labeled "relevant" (1) and "not relevant" (0) based on domain knowledge applied to their content. In this paper we use term frequency ("tf") for text content; the approach applies to more general feature vectors. Train a classifier C using a logistic regression model. Use this classifier to supervise the graph exploration process, starting from a set of relevant seed vertices. Only include in our graph those neighbors that are classified as 1 by C.

We will introduce our approach to content- and network-based graph sampling by describing two graph sampling algorithms. The first, and simplest, graph sampling algorithm (Algorithm 1) uses content only of the local vertex for classification. The second graph sampling algorithm (Algorithm 2) incorporates network effects by including the content of the neighboring nodes. The approaches are illustrated in Figure 2. These algorithms create a graph G that can be analyzed using tailored community detection algorithms [7–10, 12].

This integrated sampling and classification approach also goes hand-in-hand with network detection algorithms that rely upon weighted edges. For example, the unthresholded classifier outputs may be retained for use as edge weights [as in Eq. (1)]. Additionally, Algorithms 1 and 2 may be modified in a straightforward way to use network detection approaches for graph sampling procedure, thereby performing the steps of sampling, classification, and detection in a combined fashion.

Algorithm 1 Content-based Sampling

```
Require: seed node $s$, classifier $C$, nhop=$k$
$G = \{\}$
function EXPLORE($s$)
    for $v \in N(s)$ do
        Compute feature vector $\mathrm{tf}_v$
        if $C(\mathrm{tf}_v) == 1$ and nhop $\leq k$ then
            $G = G + v$
            Explore($v$)
        end if
    end for
end function
Return G
```

Algorithm 2 Content+Network-based Sampling

```
Require: seed node $s$, classifier $C$, nhop=$k$
$G = \{\}$
function EXPLORE($s$)
    for $v \in N(s)$ do
        $timeline_v$ = " "
        for $u \in$ random subset of $N(v)$ do
            $timeline_v = timeline_v + timeline_u$
        end for
        Compute feature vector $\mathrm{tf}_v$ using $timeline_v$
        if $C(\mathrm{tf}_v) == 1$ and nhop $\leq k$ then
            $G = G + v$
            Explore($v$)
        end if
    end for
end function
Return G
```

4. PERFORMANCE RESULTS

This section describes the raw data collection and performance results, measured using precision-recall.

4.1. Data description

To collect a real-world homophily network containing both related and unrelated content, publicly available social media content was analyzed from an active online network that shares and discusses a lexicographically distinctive subject. Seventy (70) Twitter social media users proficient in the area of cybersecurity were identified by the authors, then the public Twitter API [13] was used to analyze network content associated with these users. A $2\frac{1}{2}$ hop network using these 70 accounts as seed accounts was developed by adding new accounts that follow or are followed by source accounts, accounts from so-called @user mentions within text content, and accounts from retweet mentions. The final "1/2" hop exists because the public API is not used to access the (large quantity) of content from all known users after the second hop. Using this approach a mixed network comprised of about 34 million known accounts and over 100 million interactions was developed, with content available from 26 thousand of these accounts in the form of 34 million tweets. A small fraction of this network is illustrated using the 1-hop network shown in Figure 3.

Fig. 3. Twitter graph 1-hop network from original 70 social media accounts. Following/follower links are shown in green, @user mentions are shown in red, and retweets are shown in blue. Only a small fraction of this large graph is relevant.

As expected, much of this content is unrelated to the subject of cybersecurity, the area used to seed the network, and the objective is to classify content that is relevant to the original subject. The word clouds shown in Figure 4 illustrate the relatively different subject matter of the content between seed cybersecurity accounts and all accounts.

4.2. Performance

Figure 5 illustrates the precision-recall performance results of the algorithm of Section 3 to the data described above. The training set is comprised of 457 Twitter account labeled by the authors with 127 positive examples and 330 negative examples. Testing was performed on an independent labeled set comprised of 26 positive and 57 negative examples ($N = 83$). To generate a precision-recall curve we sweep through a classifier threshold between zero and unity.

Classifier performance improves if the content of neighboring vertices is included in the feature vector. Precision-Recall performance improvements of 20–30% are demonstrated with this dataset (green and purple curves using ranges of 1–10/and 10–50 neighboring timelines, operating point of fixed 70% pre-

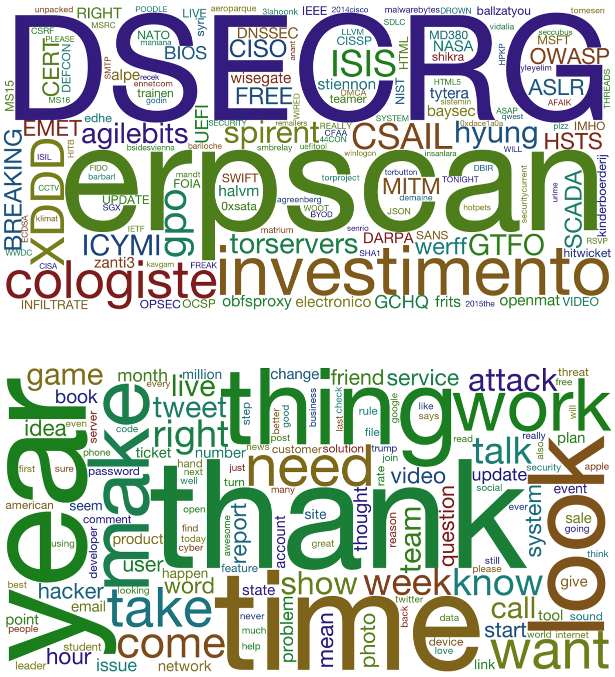

Fig. 4. Word clouds of 70 cybersecurity accounts (above) and 700 random accounts (below). Cybersecurity accounts use specific technical jargon (e.g. “torservers,” “obfsproxy,” “DNSSEC,” “IEEE”), whereas the random accounts use generic language.

cision), compared to a baseline approach that does not exploit homophily (dashed blue curve with no neighboring timelines). This is indication that homophily network effects can help bias the sampling process to the most relevant portions of the graph. Figure 5 shows that there is a threshold of diminishing returns beyond which adding more neighbor content does not lead to improved classification performance. While the threshold itself is data specific, we expect this property to hold in general settings, with the practical benefit that only a fraction of neighbors are required for prioritization. These results provide more evidence that online tight-knit affiliation communities consist of only a very small fraction of declared online friendships/ connections.

5. CONCLUSIONS

A new approach for targeted graph sampling is proposed to address the explosion of noise in online data collection systems. The algorithm uses the task of vertex classification to drive the graph sampling process towards a sample that is the most useful for the task. The algorithm captures notions of relevancy by training a classifier on a priori labeled data and exploiting the notion of content-based homophily to achieve improved classification performance. Even though we demonstrate results with online content generated features, the approach can be generalized to any vertex-based features. The performance of the algorithm is measured with real network data and content observed on a social media site. Precision-Recall performance improvements of 20–30% are demonstrated with this dataset, compared to a baseline approach that does not exploit homophily. Because real-world graphs grow exponentially, this performance improvement may have a significant impact on graph analytic algorithms with sensitivities to the graph sampling quality. As part of future work, we plan to extend the approach to consider additional content features, such as image and video based, as well as retraining the classifier at intermediate steps with the goal of improving classification performance.

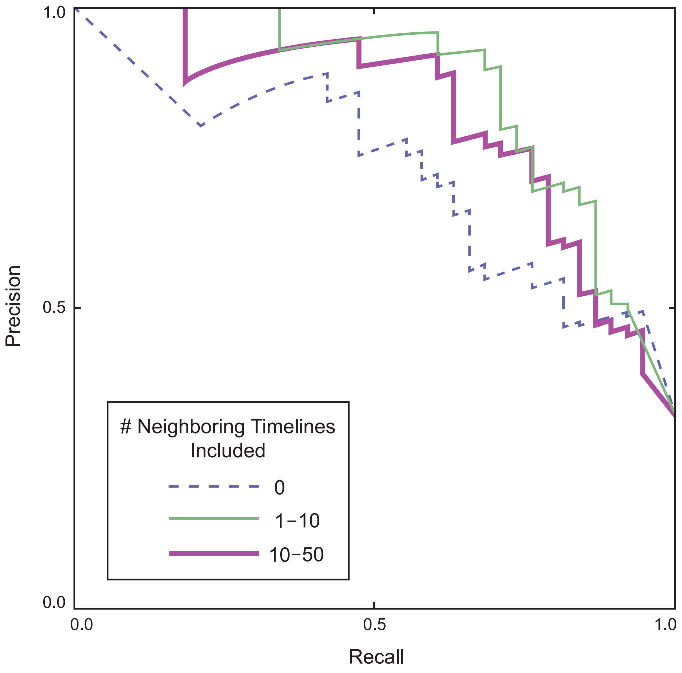

Fig. 5. Precision-Recall performance using a content and network-based classifier.

6. REFERENCES

[1] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. “Analysis of Topological Characteristics of Huge Online Social Networking Services,” in *Proc. WWW* (2007).

[2] Meng Fang, Jie Yin, and Xingquan Zhu. “Active Exploration for Large Graphs,” *Data Min. Knowl. Discov.,* **30** (3) : 511–549 (2016).

[3] J. Leskovec and C. Faloutsos, “Sampling from large graphs,” in *Proc. ACM SIGKDD* (2006).

[4] M. Gjoka, M. Kurant, C. Butts, A. Markopoulou. “Walking in Facebook: A case study of unbiased sampling of OSNS,” in *Proc. 29th Conf. Computer Communications,* San Diego CA, pp. 1–9 (2010).

[5] S. Ye, J. Lang, F. Wu. “Crawling online social graphs,” in *Proc. 12th Intl. Asia-Pacific Web Conference,* Busan pp. 236–242 (2010).

[6] A. Maiya and T. Berger-Wolf. “Online sampling of high centrality individuals in social networks,” in *Proc. 14th Conf. PAKDD,* Hyderabad, India, pp. 91–98 (2010).

[7] M. W. Mahoney, L. Orecchia, and N. K. Vishnoi, “A local spectral method for graphs: With applications to improving graph partitions and exploring data graphs locally,” *J. Mach. Learn. Res.* **13** (1) : 2339-2365 (2012).

[8] Benjamin A. Miller, Stephen Kelley, Rajmonda S. Caceres, Steven T. Smith. “Residuals-Based Subgraph Detection with Cue Vertices,” In *Proc. 49th Asilomar Conf. Signals, Systems, Computers*, 2015.

[9] M. E. J. Newman. “Finding community structure in networks using the eigenvectors of matrices,” *Phys. Rev. E* **74** (3) : 8577–8582 (2006).

[10] Tiago P. Peixoto. “Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models,” *Phys. Rev. E* **89** : 012804 (2014).

[11] Özgür Şimşek and David Jensen. “Navigating networks by using homophily and degree” *PNAS* **105** (35) : 12758–12762 (2008).

[12] S. T. Smith, E. K. Kao, K. D. Senne, G. Bernstein, and S. Philips. “Bayesian Discovery of Threat Networks,” *IEEE Trans. Signal Proc.* **62** (20) : 5324–5338 (2014).

[13] Twitter, Inc. “API Overview.” [Online]. Accessed July 2016. Available: ⟨https://dev.twitter.com/overview/api⟩.

[14] X. Zhu, J. Lafferty, and Z. Ghahramani. “Combining active learning and semi-supervised learning using gaussian fields and harmonic functions,” in *Proc. 2003 ICML Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining,* Washington DC, pp. 58–65, (2003).

[15] R. I. M. Dunbar. “Neocortex size as a constraint on group size in primates,” *J. Human Evolution* **22** (6) : 469–493 (1992).