Proceedings of 2014 INFORMS Workshop on Data Mining and Analytics (DMA 2014)
D. Sundaramoorthi, H. Yang, eds.

DEFINING "NORMAL": METRICS FOR MINING HETEROGENEOUS GRAPHS AT LARGE SCALES

Sarah Powers* and Sreenivas R. Sukumar$^{\pm}$
***Computer Science and Mathematics Division**
$^{\pm}$Computational Sciences and Engineering Division
Oak Ridge National Laboratory
Oak Ridge, TN
powersss@ornl.gov

Abstract

Rapidly increasing quantities of data are driving data science research in numerous application areas, while the ability to discover or predict patterns influences many business decisions. Large datasets are often transformed into graph-like structures to facilitate information mining. Many techniques have been proposed for studying homogenous graphs, but are ill-defined or non-existent for heterogeneous graphs, which are more suited for representing real-life datasets. This paper proposes new metrics and descriptive statistics specifically designed for large, real-world graphs with multiple vertex and edge types. To be meaningful in today's "big data" environment, new methods need to be computationally fast, memory efficient and scalable. The metrics were implemented and tested on Cray's Urika platform providing rapid results on Terabytes (TB) of data. This work provides an approach to characterizing and understanding the underlying structure of heterogeneous graphs. A use case related to the healthcare domain is presented.

Keywords: Heterogeneous graphs, information discovery, graph metrics

Introduction

In today's information-rich society, data is plentiful and no longer represents a bottleneck. This abundance creates novel areas for research, particularly in the data analysis realm. The dilemma then faced is one of synthesizing all of the data in meaningful and informative ways, a non-trivial task whose magnitude is exemplified by the sheer number of talks [1, 2], articles [3, 4] and groups dedicated to "making sense of Big Data." A similarly difficult task is finding a single piece of knowledge hidden in a vast array of extraneous data. Many cite the lack of well-developed mathematics for these large volumes of data [3, 5]. To accomplish this in many, if not all, cases, the information can be transformed and represented in a graph-like format. These alternate representations enable one to leverage many analysis techniques from graph theory. Google's PageRank algorithm [6] is a prime example, retrieving key information for users at the click of a button. Data scientists work in a wide variety of fields from social networking to transportation to health sciences. The data emerging from these areas have different characteristics resulting in diverse graphs, both homogeneous and heterogeneous.

This manuscript has been authored by Oak Ridge National Laboratory, managed by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy. The US Government retains and the publisher, by accepting the article for publication, acknowledges that the US Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US Government purposes. This research was sponsored by the Laboratory Directed Research and Development Program of ORNL.

A wide body of research exists pertaining to analysis methods for homogenous graphs and networks, which contain single node and edge types respectively. This assumption is unrealistic for many data sets [7], which tend to be more heterogeneous with multiple types of nodes and edges. These representations are well suited to represent things such as databases containing diverse types of information. A classic example is a bibliography graph such as the DBLP collaboration network consisting of many types of "nodes" (e.g., author, paper, conference) with a variety of relationships (e.g., "written by", "co-author with", "presented at"). Heterogeneous data are more information-rich, but also more difficult to work with. Despite their ubiquitousness, most analysis methods to date are tailored for homogeneous networks and do not extend well to these cases [7-9]. Existing algorithms or approaches are not necessarily portable from homogeneous to heterogeneous situations. In link prediction for example, the authors in [9] note that more information must be taken into account and thus new methods are required to address this problem in a heterogeneous context. In topic model applications, existing models can only handle homogeneous information [10]. A survey of existing analysis metrics or methods to explore and analyze heterogeneous graphs indicates a wide span of relatively recent research in the areas of anomaly detection, recommender systems, classification, prediction and topic models. With respect to traditional graph metrics, few researchers have attempted to port or create new metrics applicable to heterogeneous graphs though an unsupervised tensor-based approach to find central network nodes and perform role based clustering is presented in [11]. In this paper, we seek to address the gap in traditional graph metrics for heterogeneous networks. Specifically, we propose the *diversity degree*, a new descriptive statistic used to quantify the heterogeneity of nodes in the graph and use this metric to define the *diversity degree distribution*.

Preliminaries

We begin by introducing some key concepts along with notation. For the purposes of this study, we are interested in the interconnection of objects (e.g., doctors, authors, tests). A graph is a mathematical concept that can be utilized to represent such relationships. The terms "graph" and "network" shall be used interchangeably throughout the discussion.

Let G = (V,E) be an unweighted graph composed of the vertex set V and the edge set E. Given m types of objects $X_1, \dots, X_m$ such that V $= \bigcup_{i=1}^{m} X_i$, if m=1, then G is a homogeneous graph and if m ≥ 2, G is called a heterogeneous graph. Let e_{ij} represent the edge between any two objects i and j. To denote the possibility of one or more nodes of each type in the graph, let $X_i = \{x_{i1}, x_{i2}, \dots, x_{in_i}\} \in$ V where $n_i \geq 1$.

Metrics

In traditional graph theory, many methods exist to evaluate and compare graphs. In practice, these are defined primarily for homogeneous structures. Examples include centrality measures (e.g., degree centrality), clustering metrics (e.g., local) and distance measures (e.g., diameter/path length). To apply these to a heterogeneous graph, an abstraction is required resulting in a loss of information. For example, one of the many centrality definitions is based on highly connected vertices. In a heterogeneous network, this is ill-defined since all connections types may not be the same. Building from traditional graph theory, we define two new metrics specifically tailored to heterogeneous graphs. We seek to capture the rich information inherent in these graphs by focusing on metrics which gather diverse pieces of information into a single metric.

In a homogeneous graph, nodes can be characterized by their degree (i.e., the number of incident edges for an undirected graph) which influences a number of graph metrics such as the degree distribution, path count, and diameter of the graph. Given the propensity of heterogeneous graphs to contain not only multiple node types but also multiple edge types, in order to compute this statistic, one must synthesize the graph resulting in a loss of information. Clearly, saying that a heterogeneous node k with multiple edge and neighbor types has degree 3 masks information, containing no information about node or edge types. To overcome this loss, we propose a slight variant named the *diversity degree* defined on a per node basis. As the name suggests, the metric takes into account the diversity in the linkage (edge) types as well as the types of nodes to which a node is linked.

Definition 1: Let the *diversity degree* dv_i of a node be the sum of the edge diversity (e_{div}) and node diversity (n_{div}) for each node in the graph G.

$$dv_i = \frac{n_{div_i}}{\#node\ types} + \frac{e_{div_i}}{\#edge\ types} \tag{1}$$

Let n_{div_i} be the number of node types to which *i* is connected (regardless of edge direction or type) and e_{div_i} be the number of edge types incident to vertex *i* (regardless of direction) such that:

$$n_{div_i} = \left| \bigcap_{\forall neighbors\ j} X_j \right| \tag{2}$$

$$e_{div_i} = \left| \bigcap_{\forall e_{ij}} type(e_{ij}) \right| \tag{3}$$

In equation (1), the denominator acts as a weighting for the numerator contributions. In practice, normalization is performed (division by 2), to keep the dv_i value between [0,1] inclusive. The simplifying assumption which takes all edges as undirected is reasonable since many relations between nodes though directed are inherently unidirectional. Alternatively this metric could be computed as the sum of the node-edge-node types for each node divided by all possibilities. The proposed approach is preferred as it does not require explicitly determining the number of total possible combinations (which could be large for big graphs) and given that counting the number of edge and node types in the graph is fairly simple. By incorporating count and type, this concept captures the relational aspect of the graph instead of the structural aspect alone as in the case of vertex degrees. Cases where dv_i values and vertex degrees look similar numerically may simply indicates a lack of diversity in the relational aspect of the graph.

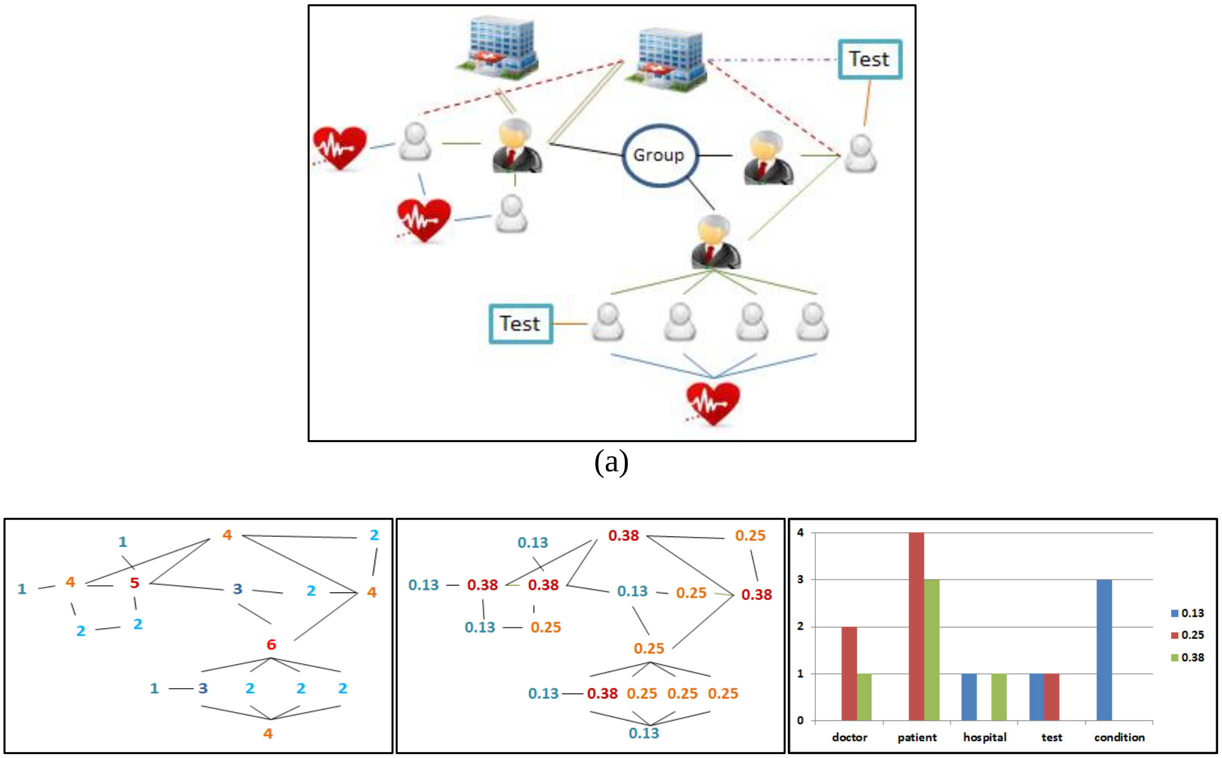
The degree distribution is a popular metric in the graph community, in part, due to the claim by Faloutsos et al. [12] that the Internet is power-law. We surmise that, similarly, there may be underlying patterns in heterogeneous graphs and propose the *diversity degree distribution* as a basis for characterizing the diversity of a network. Similar to the degree distribution, this metric is computed using the frequency of each diversity degree dv_i; the distribution is then formed by plotting the frequency versus the diversity degree.

Definition 2: Let the *diversity degree distribution* be the distribution of the diversity degrees for a heterogeneous graph G.

While clustering is often used as an important metric for quantifying community structure, we suggest that information flow can also be typified by understanding the relational structure of a network. On one hand, it has been proposed that propagation of information may occur faster and with wider spread under a diversity of connections as shown for example in the transmission of diseases (e.g., foot and mouth disease [13]). On the other hand, when searching for vulnerable spots in a network, nodes with low diversity and connectivity may be seen as weak points to protect. Using the proposed metrics, both of these types of key nodes could be determined.

Experiments and Scaling

We provide an illustrative example to highlight the metrics described. Suppose we have a healthcare provider network with 7 node types and 9 edge types (Figure 1) with associations such as "doctor prescribes test" or "patient has condition." The degrees for the "homogenized" version of the graph are shown in Figure 1(b). Figure 1(c) contains the corresponding diversity degree for each vertex. The relational structure is more apparent with some nodes rising to the forefront.



(a)

(b)

(c)

(d)

Figure 1: (a) Healthcare provider network toy example, different line types denote edge types (b) degrees of the "flattened" graph (c) diversity degrees (d) counts of the ranks by node type

Through a ranking of the diversity degrees (Fig. 1(d)), commonalities between types become more apparent. In this particular example, patients' diversity degrees are evenly distributed among options, while for the doctor category there is a possible outlier which corresponds to a doctor that does not have the highest node degree. This information is lost when the graph is flattened and this anomalous entity is not apparent (Fig. 1(b)).

To test the scalability of the metrics described, we apply these concepts on two moderate-sized data sets using Cray's *uRiKA* (Universal RDF Integration Knowledge Appliance) system [14]; a shared memory machine capable of holding entire large graphs in memory. A strength of this system is its ability to rapidly perform indexing and pattern matching. The first dataset was compiled from several health care data sets (Fred Trotter's DocGraph, the LEIE dataset of excluded providers from the Office of the Inspector General (OIG), and the NPPES database). The nodes consist of hospitals and providers, while the edges capture the relationships between them. The second dataset is a subset of the widely used DBLP collaboration network. Table 1 provides further high-level information on the datasets, as well as the time required to compute the statistics.

Table 1: Summary of datasets and computation times on uRiKA

Dataset	**Nodes**	**Edges**	**size**	**# node types**	**# edge types**	**Time (sec)**
Health Care	145278	212850	26 MB	2	18	71.23
DBLP	1068755	20263788	1.93 GB	4	19	180.83

The results indicate that for the Health Care dataset, the diversity statistic values are evenly dispersed across the lower diversity bins, with only 0.02% in the highest bin. Providers with a high diversity of associations have, in the past, been discovered to be fraudulent. Thus, this identifies key players to investigate further. The DBLP dataset has high diversity for only one node type. This is fairly intuitive as only the "paper" nodes have potential connections between all types. In general, nodes can either have one-to-one or one-to-many relations and there may not be a relationship between every class. The proposed metrics are most useful for networks where the number of node classes is less than the number of edge types. In other situations, the metrics simply highlight vertices which by definition have the most link types.

Conclusions

This paper describes two new metrics for characterizing the underlying structure of heterogeneous graphs. Many datasets are not naturally occurring graphs and require data transformation. As we continue to expand our library of datasets, these metrics will add to our fundamental understanding of the underlying graph structure. Initial computations on Cray's uRiKA indicate promising results for timely data abstraction at large scale (see [15] for additional benchmarks). This work represents a starting point, continuing with the investigation of additional datasets, identification of key commonalities, as well as integrating these new definitions into additional metrics such as centrality or assortativity.

Acknowledgments

The authors thank Tyler Brown and Gautam Ganesh for their help sanitizing and processing the data on Cray's uRiKA system.

References

1. Michel, J.-B., and Aiden, E. L., July 2011, "What we learned from 5 million books," [Video], TED, https://www.ted.com/talks/what_we_learned_from_5_million_books#.
2. Milgram, A., October 2013, "Why smart statistics are the key to fighting crime," [Video], TED, https://www.ted.com/talks/anne_milgram_why_smart_statistics_are_the_key_to_fighting_crime
3. Cold Spring Harbor Laboratory, 2014, "Better way to make sense of 'Big Data?'," ScienceDaily, www.sciencedaily.com/releases/2014/02/140218185128.htm.
4. Wolfe, P. J., 2013, "Making sense of big data," Proc. of the National Academy of Sciences, 110(25), 18031-18032.
5. Sukumar, S. R., and Ainsworth, K. C., 2014, "Pattern search in multi-structure data: a framework for the next-generation evidence-based medicine," (in preparation).
6. Page, L., Brin, S., Motwani, R., and Winograd, T., 1998, "The PageRank citation ranking: Bringing order to the web," Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, 161-172
7. Sun, Y., and Han, J., 2012, "Mining heterogeneous information networks: a structure analysis approach," SIGKDD Explorations 14(2), 20-28.
8. Lee, S., Park, S., Kahng, M., and Lee, S.-G., 2012, "Pathrank: A novel node ranking measure on a heterogeneous graph for recommender systems," Proc. of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12), Maui, Hawaii, 1637-1641.
9. Sun. Y., Barber, R., Gupta, M., Aggarwal, C., and Han, J., 2011, "Co-author relationship prediction in heterogeneous bibliographic networks," Proc. of the Advances in Social Networks Analysis and Mining (ASONAM) International Conference, July 25-27, Kaohsiung, Taiwan, 121-128.
10. Deng, H., Han, J., Zhao, B., Yu, Y., and Lin, C. X., 2011, "Probabilistic topic models with biased propagation on heterogeneous information networks," Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1271-1279.
11. Li, C.-T., and Lin, S.-D., 2012, "Centrality analysis, role-based clustering, and egocentric abstraction for heterogeneous social networks," Proc. of the 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust (SOCIALCOM_PASSAT'12), September 3-5, Amsterdam, 1-10.
12. Faloutsos, M., Faloutsos, P., and Faloutsos, C., 1999, "On Power-law Relationships of the Internet Topology," SIGCOMM Comput. Commun. Rev., 29, 251-262.
13. Sellers, R. F., and Daggupaty, S. M., 1990, "The epidemic of foot-and-mouth disease in Saskatchewan, Canada, 1951-1952," Can. J. Vet. Res., 54(4), 457-464.
14. Cray Inc., 2012, "YarcData Urika™ Enabling Real-Time Discovery in Big Data," http://www.yarcdata.com/files/product-brief/Urika%20Product%20Brief.pdf.
15. Sukumar, S. R. and Bond, N., 2013, "Mining large heterogeneous graphs using Cray's uRiKA," Proc. of the Computational Data Analytics Workshop, Oct. 8, Oak Ridge, TN.