Holistic Analysis of Multi-Source, Multi-Feature Data: Modeling and Computation Challenges

Abhishek Santra[1] and Sanjukta Bhowmick[2]

[1] Information Technology Laboratory, CSE Department
University of Texas at Arlington, Arlington, Texas, USA
[2] Department of Computer Science
University of Nebraska at Omaha, Omaha, Nebraska, USA

Abstract. As a result of our increased ability to collect data from different sources, many real-world datasets are increasingly becoming multi-featured and these features can also be of different types. Examples of such multi-feature data include different modes of interactions among people (Facebook, Twitter, LinkedIn, ...) or traffic accidents associated with diverse factors (speed, light conditions, weather, ...).
Efficiently modeling and analyzing these complex datasets to obtain actionable knowledge presents several challenges. Traditional approaches, such as using single layer networks (or monoplexes) may not be sufficient or appropriate for modeling and computation scalability. Recently, multiplexes have been proposed for the elegant handling of such data.
In this position paper, we elaborate on different types of multiplexes (homogeneous, heterogeneous and hybrid) for modeling different types of data. The benefits of this modeling in terms of ease, understanding, and usage are highlighted. However, this model brings with it a new set of challenges for its analysis. The bulk of the paper discusses these challenges and the advantages of using this approach. With the right tools, both computation and storage can be reduced in addition to accommodating scalability.

Keywords: Big Data Analytics; Multi-Source, Disparate Data; Multiplex; Graph Analysis and Query Processing; Lossless Composability; Aggregation Functions

1 Introduction

Data analytics requires a suite of various techniques to analyze different kinds of datasets and derive meaningful conclusions from them. Holistic analysis relates to analyzing a multi-feature dataset by including the effect of different combinations of features or perspectives. In this paper, we discuss a network-based model that is suited for a large class of problems. We present the utility of this model and its concomitant computing challenges.

As an example, consider the problem of modeling and analyzing the traffic accident problem or data set for a region or a country. A number of features are associated (and collected) with each accident such as location, speed, time of the day, severity of the accident, light, weather, and road conditions. One may want to analyze this dataset from multiple angles: general accident prone regions, dominant feature associated with most accidents, ordering features based on their effect on the severity of the accident, effect of individual or combination of features on accidents in a region or across all regions.

Consider another dataset where we have information about scientists who collaborate with each other, cities that have direct flights, and conferences that have overlapping research topics. In addition, there is information about who lives in which city and the cities in which annual conferences have been held. Given this dataset, it would be useful to understand: whether a large group of collaborators have attended several conferences, which group of conferences have the largest number of papers from a group of collaborators, which is the best city to hold a workshop on a particular topic to get maximum number of collaborating scientists.

Note that, unlike the earlier problem where the features referred to the same entity set (accidents), in this problem different features are captured for different disjoint entity sets (scientists, cities, and conferences). The analysis may span multiple entities and their relationships in different ways.

Traditionally, *graphs* (which we also term as monoplexes) are used for representing and analyzing systems of interacting entities [18, 21]. Typically, entities are represented as vertices. Two vertices are connected by a single edge, which represents a common value of the feature between the two entities. This representation can be extended by introducing multiple edges between vertices for each different type of feature. Instead of using multiple edges which make the representation as well as analysis of graphs difficult, we propose to use multiplexes (multiple layers of interconnected graphs) as an alternative model.

In this paper, we elaborate on the benefits of using different types of multilayer networks (homogeneous, heterogeneous, and hybrid multiplexes) for modeling and associated computation challenges for doing holistic analysis. In contrast to the vast amount of work on analyzing monoplex networks, the research on multiplexes is considerably sparse. Even when the systems are modeled as multilayer networks, they are studied only for *very specific problems in a subdiscipline* [6, 20].

We provide a brief overview of work related to multi-feature data analysis in Section 2. We will discuss modeling benefits and issues in Section 3 and the computational issues in Section 4. In Section 5, we give an overview of our preliminary work that addresses some of the challenges highlighted in this paper. We will end with the conclusions in Section 6.

2 Related Work

Recently, many analytical tasks have used multilayer networks for partitioning the space of *well-defined explicit interactions* among the *same entity set* [7,15,16, 22,25,27]. Most of the work have tried to figure out overall multiplex diagnostics such as degrees and distances by considering the multiplex layers individually or all of them together. In contrast, we focus on different types of features and entity sets and efficient analysis of arbitrary combinations of multiple layers.

Tensor Representations have also been used for certain multi-feature data representation [13]. They are mainly used for *node-aligned networks*, that is networks having same set of nodes. We are dealing with networks that are both node-aligned (homogeneous multiplex) and not node-aligned (heterogeneous multiplexes).

Graph mining (e.g., substructure discovery [14,17,23], AGM [4], FSG [14], or pattern-growth - gSpan [33], FFSM [19] and GASTON [28], disk-based approaches [5,30] and SQL-based approaches [10,29]) has been researched extensively as compared to graph querying [12]. To the best of our knowledge, graph mining and querying techniques have not been much studied for multiplexes.

3 Modeling Using Multiplexes

Multi-feature data comprises of multiple relations existing among the same or different types of entities. Relationships among the entities can either be specified by explicit interactions (like flights, co-authors and friends) or based on a similarity metric depending on the type of the feature like nominal, numeric, time, date, latitude-longitude values, text, audio, video or image.

For each feature, monoplexes will represent the relationship through directed/undirected (denoting information flow) and weighted/unweighted (quantifying relationship strength) edges between the entities, denoted by nodes. However, such monoplexes have to be generated for every feature or combination of features by repeatedly scanning the datasets and evaluating the similarity metrics. Another alternative is to use multiple edges between nodes corresponding to the features they are related to. But, in this model, for any k-feature based analysis, the entire graph will have to be loaded and traversed in order to first extract the set of desired edges. Further, such a convoluted representation makes the visualization process tedious.

In order to address the drawbacks of monoplex-based modeling, in this paper we propose the use of multiplexes, a form of *network of networks.* In this case, every layer represents a distinct relationship among entities with respect to a single feature. The sets of entities across layers, which may or may not be of the same type, can be related to each other too. Formally, a multiplex is defined by a set of n graphs $G_1(V_1, E_1)$, $G_2(V_2, E_2)$, $\ldots$, $G_n(V_n, E_n)$ and a set of edges $E_{1|2}, E_{2|3}, \ldots, E_{n-1|n}$. Each graph G_i is formed of the vertex set V_i, and the intra-layer edge set E_i. The inter-layer edge set $E_{i|j}$, connects the vertices of G_i to the vertices of G_j. Therefore, in contrast to monoplexes, for holistic analysis, the pre-processing cost is significantly reduced as the desired individual layers are either readily available or multi-feature composed layers can be generated by combining the edges of the individual layers through cost-effective set operations. Development of efficient lossless techniques for combining k individual layers translating to a new composed perspective is challenging due to the variation in edge connectivity, edge weight domain and edge directions in each layer.

Based on the type of relationships and entities, multiplexes can be of different types. Layers of a **homogeneous multiplex** are used to model the diverse relationships that exist among the **same type of entities** like traffic accidents

(Figure 1 (a)). Therefore, $V_1 = V_2 = \ldots = V_n$ and inter-layer edge sets are empty as no relations across layers are necessary. Relationships among **different types of entities** like cities (connected by flights), scientists (connected to collaborators) and conferences (related by overlapping research domains) are modeled through **heterogeneous multiplex** (Figure 1 (b)). The inter-layer edges represent the relationship across layers like conference venues, scientist residences and conference attendance. In addition to being collaborators, scientists may be friends on Facebook or connected on ResearchGate or LinkedIn. Thus, for modeling multi-feature data that capture **multiple relationships within and across different types of entity sets**, a combination of homogeneous and heterogeneous multiplexes can be used, called **hybrid multiplexes**.

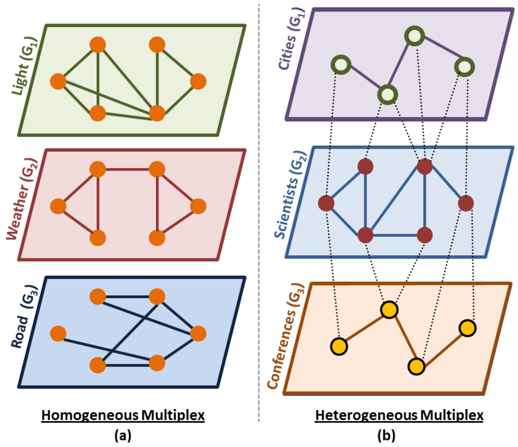

Fig. 1. Basic Types of Multiplexes

Benefits of Multiplex-based Modeling: Modeling of multi-feature data as multiplexes allows *ease of handling the dataset incrementally* through the addition of nodes (when a new accident, scientist, city or conference is encountered), edges (to represent the new entity's relationships with the earlier entities) or layers (to account for fresh perspectives). Moreover, a *latest snapshot* of multiplex can be easily maintained through the deletion of obsolete entities (nodes), relationships (edges) or perspectives (layers). Further, this modeling provides a medium to **study the relationships among the entities with respect to individual or combinations of features or perspectives**.

4 Multi-Feature Computations using Multiplexes

The major task is to be able to perform computations on the multi-feature data for holistic understanding. A plethora of algorithms are available for analyzing monoplex-based models. However, the limitations highlighted in modeling multi-feature data as a monoplex makes this medium unfavorable. On the other hand, the amount of work done for efficiently analyzing different types of multiplexes is at a nascent stage. For instance, there is hardly any work pertaining to mining and querying of multiplexes.

The traditional computational techniques proposed for monoplexes can be leveraged to perform analysis of multiplexes with respect to any combination of features (or layers). However, for holistic understanding of multi-feature data with a multiplex with n layers, $2^n - 1$ layer combinations need to be analyzed. Thus the major issue is the exponential increase in the overall computational costs with respect to both time and storage space in the presence of large number of layers ([7] has used 300 layers). This challenge highlights that the need of the hour is the **development of robust algorithms that are able to compute network characteristics, mine interesting hidden patterns and query different combinations of multiplex layers in a cost effective manner**. Additional challenges to perform specific computations on the two basic types of multiplexes have been discussed in the following sections.

4.1 Homogeneous Multiplex Computations

For the traffic accident scenario, the effectiveness of accident prevention measures and the dominance of factors can be studied through the variation in the accident-prone regions over time. In graph terminology, it translates to finding out groups of tightly connected vertices called communities (through random walks [8], maximizing modularity [26] or maximimizing permanence [9]). Therefore, for such computations we need to **devise efficient techniques for generating communities with respect to any combination of multiplex layers**. Similarly, it will be beneficial to **develop methods to compute the relative ordering and correlation among different feature (or layer) combinations based on their importance**. For example, if road conditions have *more impact* on accidents than light, then more funds can be allocated to fix the roads as compared to lights. An added challenge in this regard will be to **identify metrics that can quantify the importance of a layer** based on semantics of the domain. Density, number of influential nodes (high closeness and high betweenness centrality vertices), core-periphery structure and local and global clustering coefficients are few alternatives for such a metric.

Any of the above techniques should be efficient enough to be able to reduce the exponential complexity of generating, storing and analyzing every layer combination. **Formulation of efficient aggregation functions that can combine the results from n individual layers to compute the results of any layer combination** is a way forward. The **layer-wise analysis results will be in diverse formats** like substructures (communities), real numbers (density, clustering coefficients) or sets (hubs, high centrality nodes, nodes in inner core), adding to the complexity of this challenge. Further, the **performance of different types of network structures** for the formulated functions needs to be understood using evaluation metrics like NMI, Purity, ARI and Jaccard Index [24]. Moreover, it should be noted that there may be a class of computations for which the result of the combined layer cannot be re-constructed from the layer-wise results. For such cases, **obtaining a confidence interval for aggregation functions** will be useful to approximate results of the layer combinations.

4.2 Heterogeneous Multiplex Computations

In single networks (monoplexes) important vertices have been defined with respect to information flow through high degree, betweenness and closeness centrality vertices. However, in the case of heterogeneous multiplexes apart from the intra-layer connectivity, the inter-layer connectivity also needs to be considered.

Therefore, in the city-scientist multiplex (extracted from Figure 1 (b)), important cities will be the ones that are not only easily accessible but also where most sought after collaborators reside (marked in red in Figure 2). Thus, the challenge in this case is to **devise efficient ways to compute high centrality vertices across multiple connected layers**. It should be noted that **a high centrality vertex in one layer, may not also be a high centrality vertex in the combined layer**.

In heterogeneous multiplexes, the formulation of aggregation functions that combine the layer-wise results becomes more challenging as the **results of the bipartite graph formed by the inter-layer edges** also have to be taken into account. One must consider that **the layers may be connected not just sequentially one after another** (i.e. layer A connected to layer B connected to layer C) but can be **connected in different directions** (i.e. layers A, B and C can be connected to each other in a triangle).

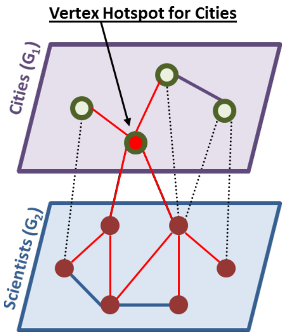

Fig. 2. Vertex hotspot for cities with respect to scientists

In monoplex networks, mining of interesting substructures of different sizes using metrics like Minimum Description Length (MDL) [17] or frequency is a well-explored field. However, to **develop algorithms for mining on multiplexes** the notion of **subgraphs and patterns** in a multiplex needs to be articulated for using a metric like *MDL* and the **anti-monotonic property of metrics** like *frequency* has to be established. The city-scientist multiplex in Figure 3 with scientist node labels depicting research fields, illustrates an example of a *frequent pattern in a multiplex*. Another challenge will be **defining exact and similar (or inexact) substructures**. Further, **strategies to partition a multiplex need to be devised** for extending the existing scalable mining techniques based on graph partitioning and map/reduce [11].

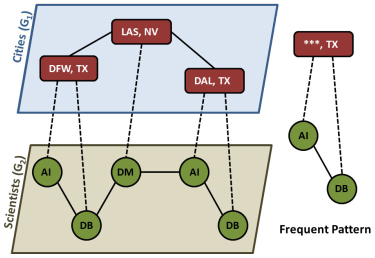

Fig. 3. Example of Frequent Pattern in a Multiplex

Querying is for verifying the existence of known patterns or extracting all instances of partially specified patterns. Queries can be of different types, for example - cities where scientists attending *most number of conferences* reside (node degree based), cities where scientists belonging to *largest group of collaborators* reside (community based), best possible city where a well-connected group of collaborators can meet up by taking the *minimum number of flights* (path based) etc. For such type of analysis, **query processing algorithms for queries on multiplexes have to be developed**. Few challenges in coming up with these algorithms are - determining the order (in parallel or as a partial order) to process layers for efficiency, generating metric to evaluate alternate query plans, evaluating the suitability of an index-based or substructure expansion-based approach and identifying query processing requirements in terms of the graph properties.

5 Preliminary Work

In this section, we will provide an overview of our preliminary work that addresses some of the challenges highlighted in this paper.

In [31], we have proposed the combination of undirected and unweighted homogeneous multiplex layers using the Boolean operators - AND, OR, NOT. For example, the AND-composed layer consists of only those edges (or relationships) that are present in all the constituent individual layers. This work proposes an intersection based aggregation method that just uses the layer-wise communities to accurately re-create the communities of any AND-composed multiplex layer, provided the communities of the individual layers are self-preserving in nature. We have shown empirically using real-life multi-feature datasets (traffic accidents [2] and storms [3]) that this community re-creation process leads to an overall saving of over 40% in computation time. Currently, we are extending this AND re-creation process to handle any type of layer-wise communities. Moreover, we are also addressing the various challenges like merging or splitting of communities based on the extent of their overlap across layers in order to formulate the community re-construction method for OR-composed multiplex layers. Metrics like modified normalized mutual information (modified-NMI) [24] that consider network topology are being used for evaluating the quality of the re-constructed communities.

Apart from communities, another recent work of ours [32] concentrates on efficiently estimating the central (or influential) entities or hubs across AND-composed homogeneous multiplex layers by using the layer-wise centrality results. Variation in the edge connectivity across individual layers can cause non-hubs to become hubs and hubs to become non-hubs in the AND-composed layers, thus making the hub estimation process a non-trivial task. Here we have developed various efficient heuristics based on degree and closeness centrality metrics by maintaining minimal neighborhood information from the individual layers. Experiments on diverse real-life multi-feature datasets (traffic accidents [2] and IMDb [1]) have shown that the proposed heuristics estimate more than 70-80% of the central vertices while reducing the overall computational time by at least 30%. Currently, we are in the process of generalizing and extending this work to other centrality measures like betweenness and eigenvector and combination methods involving disjunction (OR) and negation (NOT).

6 Conclusions

In this position paper, we have discussed the relevance of multiplexes for modeling multi-feature data as well as the computational advantages. Holistically analyzing multi-feature data can benefit from a representation that is easy to understand, visualize, and at the same time has advantages from a computation perspective.

The computational challenges identified in this paper are being addressed by us [31, 32] and the larger research community. Solutions to these challenges will enrich the data analytics repertoire making it easier to analyze problems that can benefit from graph-based representation.

Acknowledgment

We would like to extend our gratitude towards Dr. Sharma Chakravarthy, University of Texas whose insight and expertise greatly helped in shaping up of this position paper.

References

1. The internet movie database. ftp://ftp.fu-berlin.de/pub/misc/movies/database/
2. Road safety - accidents 2014. https://data.gov.uk/dataset/road-accidents-safety-data/resource/1ae84544-6b06-425d-ad62-c85716a80022
3. Storm events database by noaa. https://www.ncdc.noaa.gov/stormevents/ftp.jsp
4. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Very Large Data Bases. pp. 487–499 (1994)
5. Alexaki, S., Christophides, V., Karvounarakis, G., Plexousakis, D.: On Storing Voluminous RDF Descriptions: The Case of Web Portal Catalogs. In: International Workshop on the Web and Databases. pp. 43–48 (2001)
6. Berenstein, A., Magarinos, M.P., Chernomoretz, A., Aguero, F.: A multilayer network approach for guiding drug repositioning in neglected diseases. PLOS (2016)
7. Boden, B., Gnnemann, S., Hoffmann, H., Seidl, T.: Mining coherent subgraphs in multi-layer graphs with edge labels. In: Proc. of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2012), Beijing, China. pp. 1258–1266 (2012)
8. Bohlin, L., Edler, D., Lancichinei, A., Rosvall, M.: Community detection and visualization of networks with the map equation framework (2014), http://www.mapequation.org/assets/publications/mapequationtutorial.pdf
9. Chakraborty, T., Srinivasan, S., Ganguly, N., Mukherjee, A., Bhowmick, S.: Permanence and Community Structure in Complex Networks. (2015), accepted to TKDD
10. Chakravarthy, S., Pradhan, S.: DB-FSG: An SQL-Based Approach for Frequent Subgraph Mining. In: DEXA. pp. 684–692 (2008)
11. Das, S., Chakravarthy, S.: Partition and conquer: Map/reduce way of substructure discovery. In: International Conference on Big Data Analytics and Knowledge Discovery. pp. 365–378. Springer (2015)
12. Das, S., Goyal, A., Chakravarthy, S.: Plan before you execute: A cost-based query optimizer for attributed graph databases. In: International Conference on Big Data Analytics and Knowledge Discovery. pp. 314–328. Springer (2016)
13. De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M.A., Gómez, S., Arenas, A.: Mathematical formulation of multilayer networks. Physical Review X 3(4), 041022 (2013)

14. Deshpande, M., Kuramochi, M., Karypis, G.: Frequent Sub-Structure-Based Approaches for Classifying Chemical Compounds. In: IEEE International Conference on Data Mining. pp. 35–42 (2003)
15. Domenico, M.D., Nicosia, V., Arenas, A., Latora, V.: Layer aggregation and reducibility of multilayer interconnected networks. CoRR abs/1405.0425 (2014), `http://arxiv.org/abs/1405.0425`
16. Dong, X., Frossard, P., Vandergheynst, P., Nefedov, N.: Clustering with multilayer graphs: A spectral perspective. CoRR abs/1106.2233 (2011), `http://dblp.uni-trier.de/db/journals/corr/corr1106.html#abs-1106-2233`
17. Holder, L.B., Cook, D.J., Djoko, S.: Substucture Discovery in the SUBDUE System. In: Knowledge Discovery and Data Mining. pp. 169–180 (1994)
18. Horvath, S., Zhang, B., Carlson, M., Lu, K., Zhu, S., Felciano, R., Laurance, M., Zhao, W., Qi, S., Chen, Z., et al.: Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a molecular target. Proceedings of the National Academy of Sciences 103(46), 17402–17407 (2006)
19. Huan, J., Wang, W., Prins, J.: Efficient Mining of Frequent Subgraphs in the Presence of Isomorphism. pp. 549–552. ICDM '03, Washington, DC, USA (2003)
20. Huang, C.Y., Wen, T.H.: A multilayer epidemic simulation framework integrating geographic information system with traveling networks. In: Intelligent Control and Automation (WCICA), 2010 8th World Congress on. pp. 2002–2007 (July 2010)
21. Jeong, H., Mason, S.P., Barabási, A.L., Oltvai, Z.N.: Lethality and centrality in protein networks. Nature 411(6833), 41–42 (2001)
22. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. CoRR abs/1309.7233 (2013), `http://arxiv.org/abs/1309.7233`
23. Kuramochi, M., Karypis, G.: Frequent Subgraph Discovery. In: IEEE International Conference on Data Mining. pp. 313–320 (2001)
24. Labatut, V.: Generalized measures for the evaluation of community detection methods. CoRR abs/1303.5441 (2013)
25. Magnani, M., Rossi, L.: Formation of multiple networks. In: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. pp. 257–264. Springer (2013)
26. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69(026113) (2004)
27. Ng, M.K.P., Li, X., Ye, Y.: Multirank: co-ranking for objects and relations in multi-relational data. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1217–1225. ACM (2011)
28. Nijssen, S., Kok, J.N.: A quickstart in frequent structure mining can make a difference. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 647–652. KDD '04, ACM, New York, NY, USA (2004)
29. Padmanabhan, S., Chakravarthy, S.: HDB-Subdue: A Scalable Approach to Graph Mining. In: DaWaK. pp. 325–338 (2009)
30. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M.C.: PrefixSpan,: mining sequential patterns efficiently by prefix-projected pattern growth. In: ICDE. pp. 215–224 (2001)
31. Santra, A., Bhowmick, S., Chakravarthy, S.: Efficient community re-creation in multilayer networks using boolean operations. In: International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland. pp. 58–67 (2017), `https://doi.org/10.1016/j.procs.2017.05.246`

32. Santra, A., Bhowmick, S., Chakravarthy, S.: Hubify: Efficient estimation of central entities across multiplex layer compositions. In: 2017 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2017, New Orleans, USA, November 18, 2017. p. to appear (2017)
33. Yan, X., Han, J.: gSpan: Graph-Based Substructure Pattern Mining. In: IEEE International Conference on Data Mining. pp. 721–724 (2002)