An Introduction to Information Network Modeling Capabilities, Utilizing Graphs

Paraskevas Koukaras[1] [0000-0002-1183-9878], Dimitrios Rousidis[1] [0000-0003-0632-9731] and Christos Tjortjis*[1][0000-0001-8263-9024]

[1]The Data Mining and Analytics research group, School of Science and Technology, International Hellenic University
GR-570 01 Thermi, Thessaloniki, Greece
{p.koukaras, d.rousidis, c.tjortjis}@ihu.edu.gr

Abstract. This paper presents research on Information Network (IN) modeling using graph mining. The theoretical background along with a review of relevant literature is showcased, pertaining the concepts of IN model types, network schemas and graph measures. Ongoing research involves experimentation and evaluation on bipartite and star network schemas, generating test subjects using Social Media, Energy or Healthcare data. Our contribution is showcased by two proof-of-concept simulations we plan to extend.

Keywords: Linked data· Information networks· Graph modeling· Data mining· Social data· NoSQL.

1 Introduction

Complex Information Networks (IN) have recently drawn a lot of research interest. The structural and semantic information contained in such networks offer various new capabilities for innovation. This work reports on ongoing research including the state-of-the-art methodologies on this field of informatics and data science. Focus is given on describing aspects of complex IN, such as network schema modeling and graph measures.

Recent research addresses graph theory and Heterogeneous Information Networks (HIN) [1], both suitable for modeling multi-typed data, while exposing multiple connections and pertaining their semantic nature during any Data Mining (DM) analysis. In such structures, data objects or entities interact with many different networks, generating multilayer networks [2].

This paper presents the necessary baseline concepts for performing information modeling and testing on real or artificially generated networks, which are highly populated by multi-typed entities. For example, Social Media (SM), Energy or Healthcare/medical IN. To that end, IN analysis becomes a necessity highlighting the importance of preserving the structural integrity of these networks [3]. This type of analysis involves concepts such as network analysis, graph mining, link mining, web mining etc. [4]. The theoretical background is presented for a transition to a more practical elaboration on

IN mining. Section 2 presents the problem and approach, section 3 expands on IN modeling while referring to commonly utilized network schemas and graph measures. Section 4 presents ongoing and future research utilizing the abovementioned concepts for extracting knowledge from complex IN.

2 Problem and Approach

Most data objects, people, groups, or elements are interrelated or co-operate keeping their abstract essence. Such networks that contain non-trivial informative features are complex IN. Paradigms of IN, involve world wide web, SM, Sociopolitical, Energy, Healthcare and Academic domains while more are included in Linked Open Data Cloud[1] such as Geography, Government, Life Sciences, Linguistics etc.[4]. Semantic stores, such as OpenLink Virtuoso[2] merge RDBMS, ORDBMS, Virtual Databases etc. functionalities into a system, highlighting the importance for research on effective IN modeling and generating approaches for better data handling and knowledge extraction. Many such models and frameworks based on IN have been introduced and utilized. Despite that clustering and classification problems have benefited from IN application, just few of the proposed approaches consider network structures as they primarily focus on textual information.

This paper aims to introduce a coherent set of methods with state-of-the-art concepts that aid understanding of complex IN. It lays the foundation for the development of an approach, successfully and efficiently employing any given DM task, like forecasting [5], incorporating tools and metrics for processing data from any given domain, retrieving information and presenting results with state-of-the-art visualization tools of the fast growing graph database technologies.

One of the benefits of the proposed approach is that there are no commitments/obligations regarding the database and programming language to be used, as any NoSQL multi-model graph database can be combined with any programming language. Thus, this paper envisions the generation of two simulations, presented in section 4, for further experimentation and elaboration on complex IN. These integral parts of any graph database approach related with Big Data, in the previously mentioned domains, IN models, network schemas and graph measures providing the necessary baseline for exploiting various possibilities for innovative DM tasks. However, a wide series of tests is necessary for evaluating this theoretical approach.

3 Background

3.1 Information Network Modeling

Modeling complex IN is demanding. Real world data handling involves manipulation of data and abstract object entities that may form multilayer networks. These raw data need to be structured in a way to facilitate interactions between multiple interconnected

[1] https://lod-cloud.net/

[2] https://virtuoso.openlinksw.com/

object types, composing IN that are semi-structured. Various projects (such as KONECT[3]) address the area of network science for data collection, analysis and visualization.

Real World Networks (RWN). IN display some unique characteristics. The accuracy of the model depends on how well the model mimics real world conditions. Features amongst interconnected networks are of great importance. They associate with attributes, network analysis and statistics. In graphs networks, the degree of a node is defined as the number of the connections the node has, whilst degree distribution is the probability distribution of these degrees, scattered within the network. Clustering coefficient quantifies how close the degrees are to each other. Finally, to calculate the average path length, the average of the lengths of the shortest paths amongst all possible pairs of the network nodes is calculated [6].

Random Graphs (RG): In RG, modeling presumes that all edges connecting nodes create random relationships. This assumption though does not impose a rule and cannot always apply to real-world networks. Thus, by utilizing that model, it is assumed that all random graph relations generated always correspond to real-life networks. These networks exhibit a Poisson degree distribution, a small clustering coefficient and a normal average path [6].

Small World Model (SWM): The SWM introduces an improvement to resolve issues met within RG and more specifically issues related to the real-world representation due to issues with the clustering coefficient [7]. In SWM the average shortest path between nodes increases proportionally as a function of the number of the nodes within the network. For instance, in RWN such as SM, a person entity has a finite number of relationships (connections) like friends, groups, pages, etc. SWM approach suggests that for all entities, the number of connections is the same; therefore, all entities have the same number of neighbors. Even though SWM leads to better modeling for the clustering coefficient of RWN, there are disadvantages i.e. the unrealistic hypothesis of same number of neighboring entities and the decreased precision since the SWM produces a degree distribution similar to the Poisson degree distribution of RG.

Preferential Attachment Model (PAM). PAM seems to be the optimal and most functional model in IN modeling [8]. PAM suggests that the new nodes added to networks prefer to connect to existing ones as they share common characteristics and some already display more connections. As a node's degree increases, the probability that new nodes connect to that node increases too. Even though a PAM offers more realistic conditions (e.g. in terms of average path lengths), there are still issues with the clustering coefficient, which is very small and does not approximate the values from RWN.

3.2 Network Schemas

Multi-relational networks with single typed object. The elementary attribute of this schema is that the object type is distinct, but its relationships are always one to many. Facebook and Twitter data, along with other SM utilize this multi-relational network schema as it is more efficient for connecting, analyzing and depicting billions of links and attributes. They represent actions like messaging, sharing, connecting, publishing and many other applications [9].

[3] http://konect.cc/

Bipartite network. This schema is common in HIN and represents a relation or interaction amongst two different types of objects such as multimedia files. Bipartite networks utilize k-relations of objects creating links with other neighboring objects [10].
Star-schema network. This network schema is the most widely used conversion of relational databases where an object produces a HIN acting as a hub, where other objects connect to it. Often, relational database models such as bibliographic networks with objects of authors, books, articles etc. utilize star-schema networks [11].
Multiple-hub network. Multiple-hub networks introduce an upgrade and enhancement on star-schemas in terms of information complexity. They represent multifaceted network structures comprising many hubs, requiring increased precision in data visualization. Often, complex sciences as bioinformatics, astrophysics, theoretical mathematical structures, etc. utilize multi-hub networks where wide disintegration of network objects is required [12].

3.3 Graph Measures

Centrality's goal is to identify the central node in a graph and demonstrate the importance of vertices in graphs. Degree centrality computes a degree value defining the most central node, outlining the one with the greatest degree value. Eigenvector centrality computes the most significant node as the one with the most connections with other significant nodes. Katz centrality introduces an upgrade for the eigenvector centrality for directed graphs involving a bias term. According to betweenness centrality, the whole graph is created by multiple node hubs where the origin of these hubs are always the central nodes. Closeness centrality assumes that the nodes that proximate to the rest of the nodes are central. These measures are applied in a more common form; nodes are clustered, group degree centrality, group betweenness centrality and group closeness centrality can be distinguished [6].
Transitivity and Reciprocity. Regarding SM, manipulating the relationship between nodes (e.g. linking of nodes) is vital. Transitivity uses closed triads of edges, while reciprocity is a simpler version considering only closed loops (with length of two), that occur in directed graphs. Clustering coefficient formulas investigate the occurrences that discriminates global clustering coefficient and local clustering coefficient. These methods aid at calculating transitivity of the whole network, as well as transitivity for stand-alone nodes [6].
Similarity. Similarity is measured by referring to structural equivalence in complex IN, DM and analytics [13]. It denotes the degree to which two nodes are similar when having common neighboring nodes. An interpretation of high similarity is that nodes share the same social environments along common attributes, properties, and attitudes. Similarity levels can be computed by applying Cosine and Jaccard similarity measures [6].
Communities and Interactions. Communities can be explicit (emic) or implicit (etic) also called clusters, groups, subgroups and are vital in complex IN (e.g. Sociopolitical, Healthcare, Energy and SM DM). They involve features and dynamics that lead to the optimization of an organization's entities, representing users [14].

4 Discussion and Ongoing Research

This study showcases in an abstract way the essential theoretical background on complex IN modeling, while referring to graphs. IN modeling combined with a multi-model database, using the characteristics of a NoSQL database (e.g. Neo4J or OrientDB) [15] can offer various DM capabilities. Transitioning from SQL to NoSQL databases comes with benefits such as: 1. Import documents as with other DBMSs but also utilize relationships between objects and new data types. This is achieved by using default pointers which are persistent, enabling very fast querying, 2. Elastic linear scalability for better expanding (common master-slave architecture incommodes servers with increasing requests), 3. Open source with no limitations on development and bug reporting, 4. They support SQL querying although they are modified to work with graphs and tree structures. 5. Improved visualization with the use of graphs. 6. Enhances RDBMS capabilities by introducing concepts such as graph measures, presented in section 3.3.

To demonstrate the theoretical background of ongoing research it refers to recent literature, IN model types, network schemas and graph measures. To that end, this study prepares a baseline approach for knowledge discovery in complex IN. Contributions of this work are attributed to ongoing research, yielding two simulations for further experimentation, elaboration and result evaluation. Each simulation exposes a different notion regarding complex IN modeling, generating test subjects modeled by the concepts presented.

The *first simulation* defines a bipartite network schema while modeling and populating a database abiding with the bipartite schema and testing the validity of the model. Such a network schema displays the following characteristics: exactly two object types (nodes) with one or more relations (links) while forming a k-partite graph [10]. The dataset to be utilized refers to business reviews[4] with over 1.4 million business attributes, such as hours, parking, availability and more.

The second simulation defines a star network schema while modeling and populating a database implementing queries and calculating graph measures. Such a schema displays the following characteristics: two or more object types (nodes) with two or more relations (links) while using a HIN having the target object as a hub node. The dataset to be utilized refers to movie ratings[5] with 25 million ratings and one million tag applications applied to 62,000 movies by 162,000 users.

Current research progress attempts to inform about the key concepts that need to be considered before moving on to an effective IN analysis. Once established, future work is envisioned to involve:

A. Domain specific experimentations, such as SM, Energy or Healthcare/Medical datasets, live or historical, where user objects exist, exposing complex relations, attributes and characteristics. For example, perform sentiment analysis on SM data in comparison with identified user relationships or association rule mining or forecasting, exposing complex relationships among them.

[4] https://www.yelp.com/dataset/

[5] https://grouplens.org/datasets/movielens/

B. The incorporation of bi-functional novel algorithms like the one detailed in [16] for information extraction from very large datasets or knowledge discovery according to user specified prompts performing ranking and clustering on graphs at the same time.
C. Elaboration and evaluation of common graph measures, such as the ones presented in section 3.3, attempting to perceive new measures or metrics offering more practical applications involving user related data objects or comparing use cases with multiple graph measures.

References

[1] Han, J. (2009, October). Mining heterogeneous information networks by exploring the power of links. In *Int'l Conf. on Discovery Science* (pp. 13-30). Springer.
[2] Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2014). Multilayer networks. *Journal of complex networks*, 2(3), 203-271.
[3] Sun, Y. & Han, J. (2013). Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD, Explorations Newsletter*, 14 (2), 20–28.
[4] Koukaras, P., Tjortjis, C., & Rousidis, D. (2020). Social Media Types: introducing a data driven taxonomy. *Computing*, 102(1), 295-340.
[5] Koukaras, P., Rousidis, D., & Tjortjis, C. (2020). Forecasting and Prevention Mechanisms Using Social Media in Health Care. In *Advanced Computational Intelligence in Healthcare-7* (pp. 121-137). Springer.
[6] Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social media mining: an introduction*. Cambridge University Press.
[7] Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440-442.
[8] Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509-512.
[9] Zhong, E., Fan, W., Zhu, Y., & Yang, Q. (2013, August). Modeling the dynamics of composite social networks. In *Proc. 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining* (pp. 937-945).
[10] Long, B., Wu, X., Zhang, Z., & Yu, P. S. (2006, August). Unsupervised learning on k-partite graphs. In *Proc. 12th SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining* (pp. 317-326).
[11] Shi, C., Kong, X., Yu, P. S., Xie, S., & Wu, B. (2012, March). Relevance search in heterogeneous networks. In *Proc. 15th Int'l Conf. Extending Database Technology* (pp. 180-191).
[12] Kong, X., Cao, B., & Yu, P. S. (2013, August). Multi-label classification by mining label and instance correlations from heterogeneous information networks. In *Proc. 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining* (pp. 614-622).
[13] Koukaras, P., & Tjortjis, C. (2019). Social media analytics, types and methodology. In *Machine Learning Paradigms* (pp. 401-427). Springer.
[14] Papadopoulos, S., Kompatsiaris, Y., Vakali, A., & Spyridonos, P. (2012). Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3), 515-554.
[15] Fernandes, D., & Bernardino, J. (2018, July). Graph Databases Comparison: AllegroGraph, ArangoDB, InfiniteGraph, Neo4J, and OrientDB. In *DATA* (pp. 373-380).
[16] Koukaras, P., Berberidis C., & Tjortjis C. (2020, August). A Semi-supervised Learning Approach for Complex Information Networks. In *Proc. 3rd Int'l Conf. Intelligent Data Communication Technologies and Internet of Things* (ICICI 2020), (pp. 1-13).