ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

# Clustering Enormous and Uncertain Data Steam

A.Ashok kumar[1], Dr.P.Vivekanandan[2]

PG scholar, Dept. of. CSE, Park College of Engineering and Technology, Coimbatore-India[1]

Professor, Dept. of. CSE, Park College of Engineering and Technology, Coimbatore-India[2]

**Abstract:** Complex networks, such as biological, social, and communication networks, often entail uncertainty, and thus, can be modelled as probabilistic graphs. In a clustering problem one has to partition a set of elements into homogeneous and well-separated subsets. From a graph theoretic point of view, a cluster graph is a vertex-disjoint union of cliques. The clustering problem is the task of making fewest changes to the edge set of an input graph so that it becomes a cluster graph. Problems of Probabilistic Graph Clustering are Similar to the problem of clustering standard graphs. Probabilistic graph clustering has numerous applications, such as finding complexes in probabilistic protein-protein interaction (PPI) networks and discovering groups of users in affiliation networks. The Probabilistic Graph is generated using the Probability of occurrence of the nodes in the affiliated network. Deterministic Graph is generated depending upon the active nodes in affiliated network. The proposed system establishes a connection between our objective function and correlation clustering to propose practical approximation algorithms for Probabilistic Graph problem. A benefit of proposed approach is that the objective function is parameter-free. Therefore, the number of clusters is part of the output. The Proposed method discovers the correct number of clusters and identifies established protein relationships.

## I. INTRODUCTION

There is a growing awareness of the need for database systems to be able to handle and correctly process data with uncertainty. Conventional systems and query processing tools are built on the assumption of precise values being known for every attribute of every tuple. But any real dataset has missing values, data quality issues, rounded values and items which do not quite fit any of the intended options. Any real world measurements, such as those arising from sensor networks, have inherent uncertainty around the reported value. Other uncertainty can arise from combination of data values, such as record linkage across multiple data sources and from intermediate analysis of the data Given such motivations, Research is ongoing into how represent, manage, and process data with uncertainty.

One of the methods for finding clusters in the uncertain database is probabilistic graph clustering. This paper focuses on the problem of partitioning a probabilistic graph into clusters. The Proposed system treats every probabilistic graph G as a generative model for deterministic graphs. Each such deterministic graph is a possible world of G and is associated with a probability to be generated. Consider a deterministic graph G =(V,E) and a partitioning, C, of the nodes in V. A clustering objective function D(G,C) quantifies the cost of the clustering C with respect to G. The possible-world semantics dictate that the cost of a clustering C for a probabilistic graph G is the expected value of D(G,C), over all possible worlds of G (i.e., D(G, C) = E[D(G,C)]). Although this generalization is natural, it raises computational concerns. For instance, evaluating D(G, C) using the definition of expectation requires considering all, exponentially many, possible worlds of G. Further, the expectation of well-established clustering objective functions(e.g., the maximum cluster diameter), is infinite since, typically, there exist possible worlds where parts of the graph are disconnected. Therefore, new definitions of the clustering problem in probabilistic graphs are necessary.

We view clustering C as a cluster graph, i.e., a graph consisting of disconnected cliques. Our optimization function is the edit distance between G and cluster graph C. In other words, D(G,C) is the number of edges that we need to add and remove from G to get C. Given a probabilistic graph G, we define PCLUSTEREDIT as the problem of finding the cluster graph C that has the minimum expected edit distance from G. Our problem is a generalization of the CLUSTEREDIT problem introduced by Shamir et al for deterministic graphs.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

## II. RELATED WORK

To the best of our knowledge, we are the first to define and study the problem of clustering probabilistic graphs using the possible-worlds semantics. However, uncertain data management and graph mining has motivated many studies in the data mining and database community. We highlight some of this work here.

Graph and probabilistic-graph mining. Clustering and partitioning of deterministic graphs has been an active area of research [1]. For an extensive survey on the topic see [4] and the references therein. Most of these algorithms can be used to handle probabilistic graphs, either by considering the edge probabilities as weights, or by setting a threshold value to the probabilities of the edges and ignoring any edge with probability below this threshold. The disadvantage of the first approach is that once probabilities are interpreted as weights, then no other weights can be taken into consideration (unless the probabilities are multiplied with edge weights—in which case this composite weight has no interpretation). The disadvantage of the second approach is that there is no principled way of deciding what the right value of the threshold is. Although both the above methodologies would result in an algorithm that would output some node clustering, this algorithm, contrary to ours, would not optimize an objective defined over all possible worlds of the input probabilistic graph. Further, various graph mining problems have been studied recently assuming uncertain graphs [2].
functions between nodes in probabilistic graphs that extend shortest path distances from deterministic graphs and c proposed methods to compute them efficiently. The problem of finding shortest paths in probabilistic graphs based on transportation networks has also been considered
The intersection between the above methods and ours is that all of them deal with probabilistic graphs. However, the graph-clustering task under the possible worlds semantics has not yet been addressed by researchers in probabilistic graph mining. Data mining on uncertain data. Data mining over uncertain data has also received a lot of attention. Several classical data-mining problems have been revisited in the context of uncertainty. Examples include clustering of relational data frequent-pattern mining and evaluating spatial queries [3]. All these works are tailored to model uncertain multidimensional data where uncertainty is associated either with the location of the data points in the space or with the actual existence of the data point in the data set. One could think of defining probabilistic-graph clustering using the same ideas as those used for clustering uncertain multidimensional data. After all, the main idea there is to consider as an objective the expectation of standard clustering optimization criteria across all possible worlds [3]. It may be tempting to try and use the same definitions for probabilistic graphs, particularly since standard clustering objectives (e.g., k-center or k-median)can be optimized in deterministic graphs. However, there is a fundamental difficulty with such clustering definitions in the probabilistic-graph setting: since there are many worlds where parts of the graph are disconnected, the distance (proximity) of a node to any of the existing clustering centers can be infinity. Indeed, for nontrivial probabilistic graphs, there is always a nonzero probability of having a node with infinite distance to all the cluster centers. In that case, the optimization function becomes infinity. Therefore, new definitions of the clustering problem in probabilistic graphs are necessary. This paper addresses this challenge.

Probabilistic databases. Probabilistic databases is another active research area, mostly focusing on the development of methods for storing, managing, and querying probabilistic data There exists fundamental work on the complexity of query evaluation on such data [5], on the computation of approximate answers to queries and on efficient evaluation of top-k queries [4]. Although we borrow the possible-world semantics pioneered by the probabilistic-database community.

## III. PROPOSED METHODOLOGY

Probabilistic Graph clustering approach has problem of partitioning a probabilistic graph into clusters. This is a fundamental problem for probabilistic graphs, just it is for deterministic graphs. Partitioning a probabilistic graph into clusters has many applications such as finding complexes in protein-protein interaction networks and communities of users in social networks. A straightforward approach to clustering probabilistic graphs is to heuristically cast the probability of every edge into a weight and apply existing graph clustering algorithms on this weighted graph. This approach is

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

problematic, not only there is no meaningful way to perform such a casting, but also there is no easy way to additionally encode normal weights on the edges.

There are many algorithms are used for the cluster identification and edit distance identification in large and uncertain databases. Some of the algorithms are Agglomerative, PCast, Reference, Furthest, Balls. The performance of those algorithms is comparatively lower than the proposed system for finding Edit distance and clusters in the uncertain and large databases.[4]. The disadvantages of existing system are, Probabilities are interpreted as weights, no other weights can be taken into consideration. There is no principled way of deciding what the right value of the threshold. Non zero probability node with infinite distance to all the cluster centers. The optimization function becomes infinity.

Clustering Probabilistic Graphs has many desirable features. First Objective function can be computed in polynomial time. The value of our objective function is never infinity. We establish a connection between our problem and correlation clustering. Our algorithms also provide approximation guarantees. Partition of the nodes of the probabilistic graph into groups.

Algorithm

PKwikCluster algorithm

The quality of the solutions it produces in terms of both their edit distance and their structural characteristics compare favorably to the other algorithms.

Edit distance based

Given two deterministic graphs. We define the edit distance between G and Q, to be the number of edges that need to be added or deleted from G in order to be transformed into Q.

Results

TABLE 1
CORE Data Set: Summary of Results in Terms of Edit Distance and Wallclock Time of existing and proposed Algorithms

| Algorithm | Edit Distance | Wall clock Time |
|---|---|---|
| Agglomerative | 3420 | 10 |
| Pkwikcluster | 4192 | 0.005 |

TABLE 2
CORE Data Set: Summary of Clustering Results of existing and proposed Algorithms

| Algorithm | Cluster | Tp |
|---|---|---|
| Agglomerative | 543 | 946 |
| Pkwikcluster | 491 | 838 |

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

## IV. CONCLUSION

It presents a thorough study of clustering probabilistic graphs using the edit distance metric. The main focus is on the problem of finding the cluster graph that minimizes the expected edit distance from the input probabilistic graph. The formulation adheres to the possible- worlds semantics. Also, the objective function does not require the number of clusters as input; the optimal number of clusters is determined algorithmically. The problem here is efficiently approximated, by establishing a connection with correlation clustering. In addition, various intuitive heuristics to address it is being proposed. Further, the framework to compute deviations of a random world to the proposed clustering and to test the significance of the resulting clustering to randomized ones is established. Also, our problem where the output clustering is itself noisy is addressed. The algorithms are tested on a real probabilistic protein-protein interaction network and on a probabilistic social network that we created from Facebook! Groups. The experimental evaluation demonstrated that the algorithms not only produce meaningful clustering with respect to established ground truth, but they also discover the correct number of clusters. Finally, it is demonstrated that they scale to graphs with number of nodes and that they produce statistically significant results. number of clusters. Finally, it is demonstrated that they scale to graphs with number of nodes and that they produce statistically significant results.

## REFERENCES

[1] Ailon.N, M. Charikar, and A. Newman, "Aggregating Inconsistent Information: Ranking and Clustering," Proc. Annual ACM Symposium. Theory of Computing (STOC), pp. 684-693, 2005.
[2] Bansal.N, A. Blum, and S. Chawla, "Correlation Clustering," Machine Learning, vol. 56, nos. 1-3, pp. 89-113, 2004.
[3] Benjelloun.O, A. Halevy, J. Widom, and A.D. Sarma, "ULDBs: Databases with Uncertainty and Lineage," Proc. 32nd Int'l Conf. Very Large Data Bases (VLDB), pp. 953-964, 2006,
[4] Bonchi.F, A. Gionis, G. Kollios and M. Potamias, "K-Nearest Neighbors in Uncertain Graphs," Proc. VLDB Endowment, vol. 3, nos. 1/2, pp. 997-1008, 2010.
[5] Cheng.R, D.W. Cheung, J. Cheng and L. Sun,, "Mining Uncertain Data with Probabilistic Guarantees," Proc. 16th ACM SIGKDD International Conference Knowledge Discovery in Data Mining (KDD), pp. 273-282, 2010.
[6] Cormode.G and McGregor.A , "Approximation Algorithms for Clustering Uncertain Data," Proc. 27th ACM SIGMOD-SIGACTSIGART Symposium. Principles of Database Systems (PODS), pp. 191-200, 2008.
[7] Dalvi.N.N and D. Suciu, "Efficient Query Evaluation on Probabilistic Databases," Proc. 13th Int'l Conf. Very Large Data Bases (VLDB), 2004.
[8]Dalvi .N.N, Re.C, and Suciu.D , "Probabilistic Databases: Diamonds in the Dirt," Comm. ACM, vol. 52, no. 7, pp. 86-94, 2009.
[9] Frank.H, "Shortest Paths in Probabilistic Graphs," Operations Research, vol. 17, no. 4, pp. 583-599, 1969.
[10] Gao.H, J. Li and Z. Zou, "Discovering Frequent Sub graphs over Uncertain Graph Databases under Probabilistic Semantics," Proc. 16th ACM SIGKDD International Conference Knowledge Discovery and Data Mining (KDD), pp. 633-642, 2010. 63
[11] Krogan.N.J, "Global Landscape of Protein Complexes in the Yeast Saccharomyces Cerevisiae," Nature, vol. 440, no. 7084, pp. 637-643, http://dx.doi.org/10.1038/nature04670, Mar. 2006.
[12] Schaeffer.S.E, "Graph Clustering," Computer Science Rev., vol. 1, no. 1, pp. 27-64, 2007.
[13] Shamir.R, R. Sharan, and D. Tsur, "Cluster Graph Modification Problems," Discrete Applied Math., vol. 144, nos. 1/2, pp. 173-182, 2004.
[14] Valiant.L.G, "The Complexity of Enumeration and Reliability Problems," SIAM J. Computing, vol. 8, no. 3, pp. 410-421, 1979.