# Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 11 (17): 3221-3232, 2011
ISSN 1812-5654 / DOI: 10.3923/jas.2011.3221.3232


# Connection Subgraphs: A Survey

[1]Nouf M. Kh. AlSudairy, [2]Vijay V. Raghavan, [3]Alaaeldin M. Hafez and [1]Hassan I. Mathkour
[1]Department of Computer Science, College of Computer and Information Sciences,
King Saud University, Saudi Arabia
[2]Center for Advanced Computer Studies, University of Louisiana, USA
[3]Department of Information Systems, College of Computer and Information Sciences,
King Saud University, Saudi Arabia

**Abstract:** Mining large graphs to discover relationships between two or more nodes is an important problem. This study presents a literature review on a specific formulation of that problem which is referred to as the connection subgraph problem. Connection subgraphs are useful in many applications such as ranking search results, discovering connections between criminals or terrorists, identifying connections between two genes, exploiting product relationships to increase product sales and visualizing large graphs. The study also presents suggestions for future research directions.

**Key words:** Connection subgraph, connection subgraph problem, CSDP, n-CSDP, CEPS, proximity graphs, link mining, graph mining, link analysis, relational learning

## INTRODUCTION

Graphs are important in modeling the interaction between nodes and structures. It has many applications such as computer vision (Kandel *et al*., 2007), pattern recognition (Kandel *et al*., 2007), web analysis (Chakrabarti, 2002) and text retrieval (Shehata *et al*., 2006).

Traditional data mining techniques take homogenous objects from one relation which may not be suitable with graphs. Graph data may be heterogeneous, semi-structured and multi-relational (Cook and Holder, 2006; Aggarwal and Wang, 2010; Maimon and Rokach, 2010). Traditional data mining techniques such as classification (Sharma *et al*., 2007; Mokeddem and Belbachir, 2009; Phyu, 2009), clustering (Raghavan and Birchard, 1979; Raghavan and Yu, 1981; Noel *et al*., 2003; Berkhin, 2006; Velmurugan and Santhanam, 2011) and frequent pattern mining (Hafez *et al*., 1999; Raghavan and Hafez, 2000; Yoon and Raghavan, 2000; Kubat *et al*., 2003; Thakur *et al*., 2007; Tiwari *et al*., 2010) have been extended to work with graphs. The resulting techniques are more challenging (Aggarwal and Wang, 2010). Specifically, a new emerging research area has appeared under the name of link mining (Noel *et al*., 2003; Getoor and Diehl, 2005). Link mining refers to data mining techniques that take into account links (i.e. relationships between objects) while building descriptive or predictive models. Thus, the mining process has more available information which leads to many new tasks including object ranking, object classification, object clustering, object identification, link prediction, subgraph discovery and graph classification. Link mining is at the intersection of the work in graph mining, hypertext and Web mining, link analysis, relational learning and inductive logic programming (Getoor, 2003; Getoor and Diehl, 2005; Han and Kamber, 2006; Yu *et al*., 2010).

Mining large graphs to discover relationships between two or more nodes is an important problem. This study presents a literature review on a specific formulation of that problem which is referred to as the connection subgraph problem.

## PRELIMINARIES

A graph $G = (V, E)$ consists of a set of vertices $V = \{v_1, v_2, \ldots\}$ and a set of edges $E = \{e_1, e_2, \ldots\}$. Each edge $e_i$ is associated with a pair of vertices $\{v_j, v_k\}$, $v_j, v_k \in V$, called its end vertices. In a simple graph, edges do not hold any attributes other than their end vertices. However, in a more sophisticated graph, edges can have directions weights and/or labels. Vertices can also have weights and/or labels. $G' = (V', E')$ is a subgraph of $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$. In other words, graph $G = (V, E)$ contains graph $G' = (V', E')$ (Wallis, 2007; Fournier, 2009; Bapat, 2010).

**Corresponding Author:** Nouf M. Kh. AlSudairy, Department of Computer Science, College of Computer and Information Sciences, King Saud University, Saudi Arabia

Graphs can form many types of networks such as social networks, biological networks and computer networks. A social network is a set of vertices which are directly or indirectly related to each other depending on a common interest. The context of social networks does not have to be social. There exist several real-world instances of biological, business and technological social networks such as cellular and metabolic networks, the World Wide Web and telephone call graphs (Ahmad and Rizwan, 2005; Han and Kamber, 2006).

## PROBLEM DESCRIPTION

Using a single path to represent the relationship between two vertices is limiting because mistakes may happen when selecting the most important path. Therefore, providing a subgraph (i.e., many paths) increases the probability of having the critical path. Furthermore, there might not be a critical path. Thus, the subgraph is necessary in this case. Faloutsos *et al.* (2004a) showed that both the shortest path and network flow measures fail to select the best path in social networks. They considered a graph similar to the one shown in Fig. 1 and assumed that the weights of all edges are 1. Thus, the shortest path measure fails to distinguish between paths s, 4, t and s, 5, t, since both paths have a length of 2. However, it is clear that node 5 has many edges and hence the path through vertex 4 is preferred. Similarly, the network flow measure does not distinguish between paths s, 1, 2, 3, t and s, 4, t since both paths carry 1 unit of flow. However, it is clear that the latter path is preferred since it is shorter.

The connection subgraphs mining problem was introduced by Faloutsos *et al.* (2004a). The same formulation was applied to the context of social networks by Faloutsos *et al.* (2004b). They defined a connection subgraph as a small connected subgraph of a large graph that well captures the relationship between two non-adjacent nodes. The relationship represents a set of interesting paths between the nodes. For example, in a social network of people, the connection subgraph shows how two individuals are related. They formalized the Connection Subgraph Discovery Problem (CSDP) as follows: Given an edge-weighted undirected graph G = (V, E), two (n = 2) vertices s and t and an integer budget b. Find: a connected subgraph that contains s and t and at most b other vertices which maximizes a specified goodness function. The connection subgraph problem has two sub-problems. The first one is determining what function is appropriate for measuring goodness and the second one is devising an algorithm using which the subgraph can be found quickly. Later on, the concept of connection subgraphs was extended to work with n (>2) nodes.

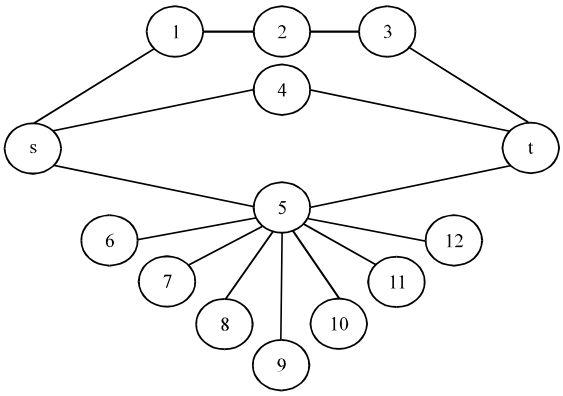


Fig. 1: Sample network

## CONNECTION SUBGRAPHS

Large graphs can be mined to discover connection subgraphs between two or more nodes. This section presents a literature review of research on connection subgraphs. The first subsection deals with finding a connection subgraph between two given nodes while the second and third subsections deals with finding connection subgraphs between multiple nodes.

**Connection subgraph between two nodes:** Faloutsos *et al.* (2004a) proposed an efficient algorithm for CSDP that is based on a model of a large social network as an electrical resister network where the delivered current between the two nodes is to be maximized. The two vertices s and t represent, respectively, the source and destination of an electrical current. The proposed goodness function is the total delivered current that the subgraph carries from s to t. The proposed algorithm discovers the paths through which current can flow. Furthermore, all nodes are connected to a universal sink node that observes a positive proportion of the current that flows into each node. Thus the universal sink node penalizes high-degree nodes as well as long paths. They reduced the problem to the following steps: first, calculating the currents on the given graph and; second, extracting the subgraph that carries more current to t. Since calculating current flows with the universal sink node in a very large graph is not possible in an interactive environment, they also proposed an optional preprocessing step which quickly reduces the size of the graph by removing nodes and edges that are far from both s and t. They have

demonstrated how to quickly generate high-quality connection subgraphs from large graphs. The main motivation for their work was to give a paradigm for knowledge discovery in large social network graphs. Figure 2 shows a connection subgraph that was generated by their interactive system. The connection subgraph shows the relationship between Alan Turing and Sharon stone. Edge weights show the relative connection strengths (Original edge weights are shown in parentheses).

The algorithm of Faloutsos *et al*. (2004a) is only appropriate for finding a connection subgraph between two nodes; thus, does not generalize for the case more than two input nodes. This is because the algorithm is based on the electrical currents in a resistor network. Another limitation of Faloutsos *et al*.'s algorithm is that, because of current loss, the resulting connection subgraph may differ depending on which of the two nodes (s or t) was selected as the source.

As shown in Fig. 2, Faloutsos *et al*. (2004a) used a graph in which all nodes represented the same concept (i.e., famous people) and therefore all edges have the exact same interpretation (i.e., strength of acquaintance between two people). The situation is different with Resource Description Framework (RDF) graphs (W3 Org, 2004).

Ramakrishnan *et al*. (2005) proposed a way to find informative subgraphs in RDF graphs. An RDF graph represents the structure of an RDF expression. An RDF graph is composed of a set of triples where each triple consists of a subject, a predicate and an object. The subject and the object of each triple are linked by a predicate (W3 Org, 2004). Therefore, the weighting scheme that Faloutsos *et al*. (2004a) used cannot be used to find relevant subgraphs in RDF graphs. Moreover, it is not sufficient to use a uniform weight on all edges because edges in RDF graphs may have different semantics. Thus, the weighting scheme needs to be based on the semantics suggested by the RDF schema. In Ramakrishnan *et al*. (2005), first, they assigned weights to all edges based on different informativeness measures. Then, they extracted the connection subgraph such that it maximizes a goodness function based on Faloutsos *et al*. (2004a) resistor network model. Figure 3 shows a connection subgraph resulting from their synthetic dataset. The connection subgraph shows the relationship between Actor-5567 and Captain-8262. Nodes are colored based on the schemas that their classes belong to. Dark grey nodes are used to represent Entertainment schema classes, Light gray nodes represent Business schema classes and white nodes represent Sports schema classes.

Sevon and Eronen (2008) presented a technique that used context-free grammar to specify the class of interesting paths that are useful for querying labeled graphs where vertices are labeled by a type from a set of vertex types (e.g., gene, phenotype, etc.) and edges are

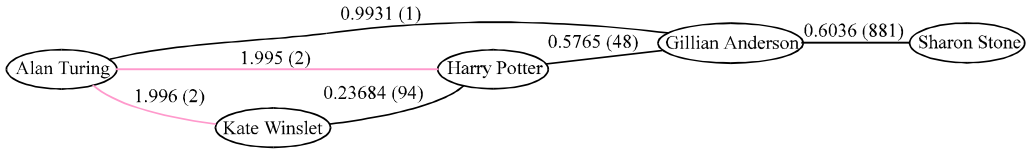


Fig. 2: A connection subgraph from Faloutsos *et al*. (2004a) interactive system

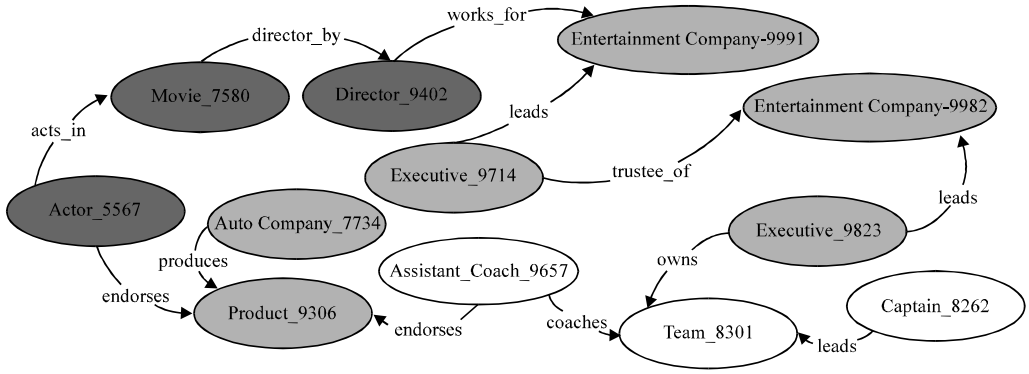


Fig. 3: A connection subgraph resulted from Ramakrishnan *et al*. (2005) synthetic dataset

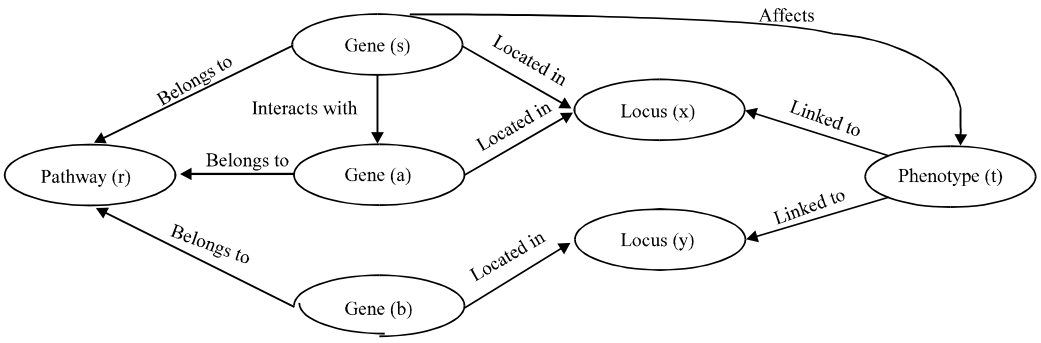


Fig. 4: An example of a labeled connection subgraph

labeled by a type from a set of edge types which describes the type of relations between vertices (e.g., belongs to, interacts with, etc.). The union of the interesting paths between two vertices, or two sets of vertices, is the resulting connection subgraph. The main motivation of this technique is that it does not produce a connection subgraph that includes all paths connecting the given two vertices. Instead, it produces a connection subgraph that only includes interesting paths with specific semantics. Labeled graphs queries are usually based on vertex types and edge types.

Figure 4 shows a similar example to that illustrated by Sevon and Eronen (2008). It shows a labeled connection subgraph capturing the interesting paths between a gene (s) and a phenotype (t). The authors also presented algorithms for generating the connection subgraph directly, i.e., without having to enumerate all the individual interesting paths. Experimental results that were obtained on a set of real graphs derived from some biomedical databases showed a superior performance.

Jin *et al.* (2007a) proposed a technique that detects links between two concepts in text documents. They presented the idea of Concept Chain Queries (CCQ) which are useful for discovering top k most meaningful evidence trails in documents that connect the two concepts. They suggested using link-analysis techniques on the extracted features in order to discover new knowledge. They have compared two approaches. The first is the Concept-profile approach that uses the bag-of-words model and the second is the Graph-based approach that combines techniques from graph mining, text mining and link analysis. In the latter approach, potential conceptual connections are discovered and concept chains are created and ranked based on the weights of the corresponding selected chains. Experimental results showed that the Graph-based approach is favored for discovering focused information. However, the Concept-profile approach is favored for greater coverage of information. The Graph-based approach is similar to the problem presented by Faloutsos *et al.* (2004a). However, there are some important differences. The proposed approach relies on URL links to discover connections between documents, addresses general concepts, generates explanations of the chains and does not discover all paths together. Instead it discovers the paths individually which permits more user input (e.g., novelty, recency, etc.,) in order to determine the best paths. Similar work is found by Srihari *et al.* (2005a,b, 2007) and Jin *et al.* (2007b). Figure 5 shows an example of a CCQ that was proposed by Jin *et al.* (2007b) where a possible connection between Nashiri and Nairobi attack was discovered through the concept Jihad Mohammad Ali al Makki.

Doc 1: **Nashiri** and his cousin, Jihad Mohammad Ali al Makki, returned to Afghanistan, probably in 1997, Nashiri again encountered Bin Ladin, still recruiting for "the coming battle with the United States." Nashiri joined al Qaeda and later was recognized as the chief of al Qaeda operations in and around the Arabian Peninsula.

Doc 2: In late 1998, al Qaeda decided mounting an attack against a U.S. vessel and Jihad Mohammad Ali al Makki, also known as Azzam was a suicide bomber for the **Nairobi attack.**

Fig. 5: Concept chain and evidence trail (Jin *et al.*, 2007b)

**Connection subgraphs between multiple nodes:** Some authors extended the connection subgraph discovery problem to the n-Connection Subgraph Discovery Problem (n-CSDP). The goal of n-CSDP is to find a small connected subgraph of a large graph that well captures the relationship among n nodes where n is greater than 2 (Chen *et al.*, 2006). Conrad *et al.* (2007) investigated the complexity of the n-CSDP. They presented results on its worst-case complexity and empirical typical-case. They

showed that the decision version of the n-CSDP is NP-complete and that both the cost and utility optimization versions are NP-hard.

Vast *et al*. (2005) proposed a technique that extracts a subgraph that will capture the relevant edges among n given nodes in a biochemical network. The technique projects the nodes of the network that is represented as an undirected graph into a Euclidean space. The proposed approach has two phases. The first phase extracts a subset of relevant nodes in the graph. The second phase constructs a subgraph that connects them. Chen *et al*. (2006) tested Vast *et al*. (2005) algorithm on two real networks. Experimental results showed that both subgraphs produced were disconnected. In addition to this problem, the time cost is unacceptable if the network is very large. Chen *et al*. (2006) also proposed an approach to solve the n-CSDP in two phases. The first phase extracts a fairly big candidate subgraph using the proposed method based on neighborhood-growth. The second phase uses an evolutionary algorithm (Eberhart and Shi, 2007) for optimizing the extracted subgraph. In order to encode the topology of subgraphs as individuals, a transformed representation of the adjacent matrix of graphs, called UTM (Upper Triangle Matrix) code, is designed. Also, some evolutionary operators are provided to be directly used on the UTM code. In order to fully benefit from the evolutionary computational algorithm, it was suggested that the size of the candidate subgraph be a number of times larger than the result subgraph. Experimental results on a real dataset showed that the algorithm is scalable and has good performance. One limitation of their algorithm is that it only considers one type of entity (e.g., people). However, removal of this constraint is stated as a possible direction for future work.

Tong and Faloutsos (2006) presented and solved the Ceenter-Piece Subgraph (CEPS) problem. The problem finds the subgraph that has strong connections to most or all of the n nodes in a social network. The connection subgraph extraction algorithm seeks to maximize the goodness score of the nodes. In order to do that, each source node simulates independently a Random Walk with Restart (RWR) (Tong *et al*., 2006; Tong *et al*., 2008a) and the goodness score of a node is calculated by the probability that the random particles are going to meet at that particular node. Next, important paths are discovered iteratively using dynamic programming. The authors also introduced and handled the K_softAND queries that are in between AND and OR queries. In K_softAND, a node is important if it is important to at least k out of Q source queries. Figure 6 shows an example of a query among four nodes.

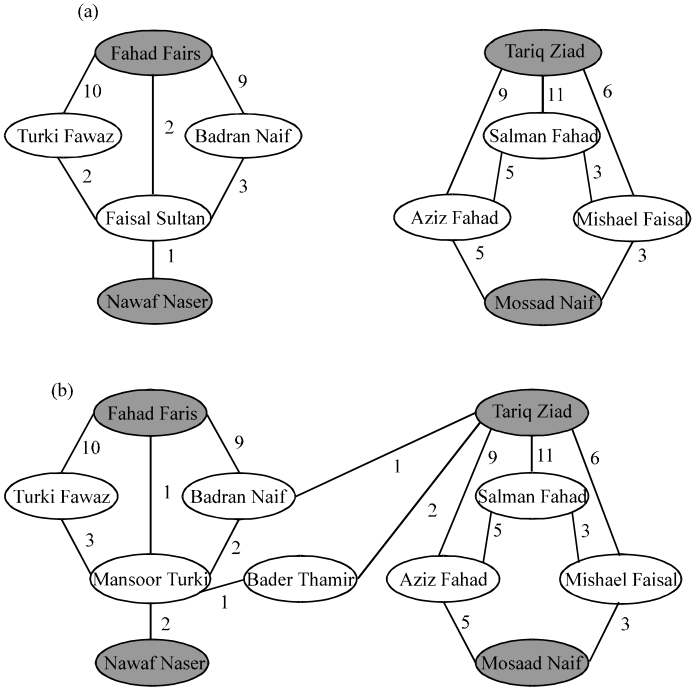


Fig. 6: CEPS among four nodes, (a) 2_softAND query and (b) AND query

In Fig. 6a, the user asked for 2_softAND query which is reasonably not connected since Fahad Faris and Nawaf Naser belong to the data mining community and Tariq Ziad and Mosaad Naif belong to database community. In contrast, in Fig. 6b, the user asked for AND query therefore the subgraph shows strong connections between all four nodes. Experimental results on a large real dataset showed that the algorithm is very fast and almost accurate (~90%) (Rodrigues *et al*., 2006; Tong and Faloutsos, 2006).

Later on, Tong *et al*. (2007a) studied the direction-aware proximity measure and generalized CEPS to work with directed graphs. The proposed proximity measure is based on the escape probability notion of random walks. The key modification that they introduced to the original CEPS is that of using the direction-aware proximity measure to calculate the goodness scores. They introduced a token vector $\vec{f} = [f_1, ..., f_Q]$ where $f_i \in (+1, -1)$. This divides the query nodes into source(s) and target(s) so that proximities and paths are discovered from source(s) to target(s). The generalized algorithm is called Dir-CEPS. Figure 7 shows an example of how Dir-CEPS may discover many relationships between the same query nodes. This is done by selecting different token vectors. Figure 7a, labeled $\vec{f} = [+1, -1]$, shows how the Paper No. 1111 influences the Paper No. 2222. Figure 7b, labeled $\vec{f} = [+1, +1]$, shows how the two papers, Paper No. 1111 and Paper No. 2222 influence other papers. Figure 7c, labeled $\vec{f} = [-1, -1]$, shows how the two papers, paper No. 1111 and paper No. 2222 are influenced by other papers. Experimental results showed that the new measure is useful for several applications and attains a significant speedup over straightforward implementations.

Connection subgraphs discovery is one way to show proximity among nodes in networks. Koren *et al*. (2006, 2007) proposed the Cycle Free Effective Conductance (CFEC) technique which is another way of measuring and extracting proximity in networks. Their work is heavily inspired by the work of Faloutsos *et al*. (2004a) and may serve the same purpose as the connection subgraphs but captured with CFEC technique. However, the goal of this work is to extract the subgraph that retains most of the proximity between the terminal nodes. The proposed technique can also show connections among more than two nodes, work with directed graphs and solve optimization problems with tunable parameters. The new proximity measure has an intuitive interpretation that is based on random walks. The proposed technique is capable of predicting connections that have not yet been

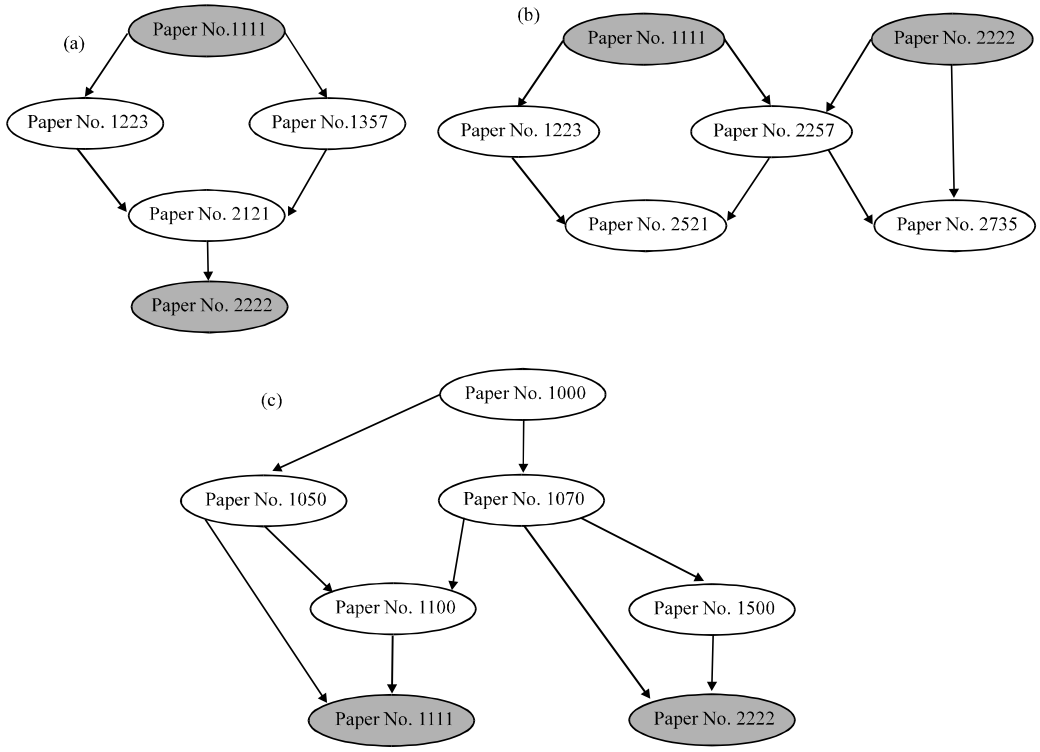


Fig. 7: Several relationships between the same query nodes, (a) $\vec{f} = [+1, -1]$, (b) $\vec{f} = [+1, +1]$ and (c) $\vec{f} = [-1, -1]$

made within social networks, determining links that cannot be observed in a network with missing data and finding communities of entities in a network that behave similarly.

Asur and Parthasarathy (2009) proposed a model that finds the ViewPoint Neighborhood (VPN) of a node or a set of nodes. They defined the VPN for a given node x as the graph that contain only the nodes that have some degree of importance to x and their interconnections. In case of having a set of source nodes, the problem is to discover the VPN that represents the intersection of the neighborhoods of at least k nodes. In order to easily extract the interesting neighborhoods, they constructed some activation functions leveraging topological and semantic information. The multi-node neighborhoods problem is similar to the CEPS problem. The main difference is that the proposed algorithm makes use of other constructs like topological properties and semantic information in order to construct neighborhoods.

**Connection subgraphs between multiple nodes using context-aware object connection discovery:** All of the above work considered the connection between individual objects. In some real world problems, objects may have a context such as a terrorist belongs to a terrorist group. Cheng *et al.* (2009) suggested a context-aware object connection discovery in a large graph. They partitioned the large graph into communities based on a modularity concept such that every community becomes the context of the nodes within the community. The proposed technique has two phases. The first phase computes the best intra-community connection. This is done by maximizing the amount of information flow. The second phase extends the connection to the inter-community level. This is done by using the community hierarchy relation. Dividing the large graph into smaller communities and considering the context has improved the efficiency of the connection discovery process. Experimental results showed that the algorithm is efficient. The suggested algorithm was compared with the CEPS. Results showed that the quality of the subgraphs generated is comparable. However, the suggested algorithm is three orders of magnitude faster than CEPS and consumes less memory. Figure 8 shows the context-aware connection subgraph for {Fahad Faris, Nawaf Naser, Tariq Ziad, Mosaad Naif}. The technique first discovers that the four people are from two contexts. Fahad Faris and Nawaf Naser are from the data mining community and Tariq Ziad and Mosaad Naif are from the database community. Then, the technique finds the best connection between the people within each context. Finally, the connections between the two contexts are found.

Table 1 presents a summary of the various classes of graph search problems and the solution approaches.

## VISUALIZATION TOOLS

There are many graph visualization tools available. The problem with them is that they have a difficulty in handling large graphs and/or they don't permit interaction. Large graphs as pointed out by Rodrigues *et al.* (2006) have two challenges. Firstly, straightforward interactive manipulation is prohibitively slow. Secondly, even if the graph can be plotted and replotted quickly, the user is going to be overwhelmed with the large volume of information. Therefore, Rodrigues *et al.* (2006) proposed the GMine system which addresses the two aforementioned challenges by using summarization and multi-resolution. GMine provides summarization by implementing the connection subgraph extraction algorithm discussed by

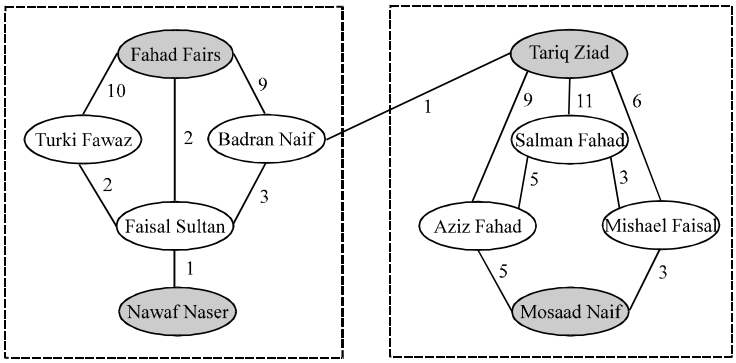


Fig. 8: Context-Aware Connection subgraph for {Fahad Faris, Nawaf Naser, Tariq Ziad, Mosaad Naif}

Table 1: Summary of different connection subgraph algorithms

| Reference | No. of Query nodes | No. of Vertex types | No. of Edge types | Description |
|---|---|---|---|---|
| Faloutsos *et al.* (2004a) | 2 | 1 | 1 | • Models a large social network as an electrical resistor<br>• Discover paths through which current can flow |
| Ramakrishnan *et al.* (2005) | 2 | n | n | • Works with RDF graphs<br>• Based on Faloutsos *et al*'s resistor network model |
| Jin *et al.* (2007a) | 2 | 1 | 1 | • Detects links between two concepts in text documents<br>• Presented the CCQs |
| Sevon and Eronen (2008) | 2 | n | n | • Works with labeled graphs<br>• Uses context free grammar<br>• Shows only interesting paths |
| Vast *et al.* (2005) | n | n | 1 | • Projects the nodes of the undirected graph into an Euclidean space<br>• Extracts a subset of relevant nodes and constructs a subgraph that connects them |
| Chen *et al.* (2006) | n | 1 | 1 | • Extracts a fairly big candidate subgraph and uses an evolutionary algorithm for optimizing it |
| Tong and Faloutsos (2006) | n | 1 | 1 | • Called CEPS<br>• Uses RWR and dynamic programming<br>• Introduced the *K_softAND* queries |
| Tong *et al.* (2007a) | n | 1 | 1 | • Generalized CEPS to Dir-CEPS in order to work with directed graph<br>• Introduced the token vector $\vec{f}$ |
| Koren *et al.* (2007) | n | 1 | 1 | • May serve the same purpose as the connection subgraphs<br>• Uses CFEC technique |
| Asur and Parthasarathy (2009) | n | 1 | 1 | • Finds VPNs<br>• Similar to CEPS but makes use of other constructs like topological properties and semantic information |
| Cheng *et al.* (2009) | n | 1 | 1 | • Uses context-aware object connection discovery<br>• Partitions the graph into communities based on the modularity concept<br>• Computes the best intra-community connection and then extends it to the inter-community level |

Tong and Faloutsos (2006) and visualizing the output. It provides multi-resolution graph exploration by partitioning the graph into a hierarchy of communities. Thus, the system speeds up large graph exploration by focusing on a subset of the graph.

Proximity graphs such as the Minimal Spanning Tree (Deo, 2004) can also be used to display a complex graph on a two-dimension plane. There are many algorithms to compute minimal spanning trees whether for directed or undirected graphs. Using a minimal spanning tree to formulate the graph layout may also be a solution in gaining layout predictability where two runs of the algorithm using the same or similar graphs does not lead to drastically different representations (Herman *et al.*, 2000). Even though spanning trees plays an important role in graph visualization, there are other layout approaches, such as Delaunay Triangulation (Bose *et al.*, 1996), Relative Neighborhood graph (Toussaint, 1980) and Gabriel graph (Gabriel and Sokal, 1969), that deserve consideration as methods for visualizing and exploring large graphs.

## RELATED WORK

A closely related concept to the connection subgraph was introduced by Hintsanen (2007). He introduced the concept of the most reliable subgraph. Both concepts have a very similar goal. However, the latter is based on probabilistic reasoning. It can be said that the most reliable subgraph problem is an instance of the connection subgraph problem. Hintsanen formulated the problem of finding the most reliable subgraph as follows: Given: a graph subject to random edge failures, a set of terminal vertices and an integer K. Find: a subgraph that contains K fewer edges than the original graph and maximizes the probability of connecting the terminal vertices. The author focused on the two-terminal and undirected graph case. Experimental results demonstrated the effectiveness of the most reliable subgraph concept. Later on, Hintsanen and Toivonen (2008) suggested two new heuristics for solving the most reliable subgraph problem. Best Paths Incremental (BPI) and Series-Parallel Augmentation (SPA). They provided technical details as well as a rough complexity analysis of the suggested algorithms. Experimental results showed that the algorithms are scalable.

Sevon *et al.* (2006) presented a way to discover links in biological databases by proposing a technique for measuring the link's strength among two vertices. The proposed measures depend on the following issues: reliability, relevance and rarity. They presented techniques that find good paths and subgraphs in addition to evaluating their quality. These techniques are not only applicable to gene-phenotype links but also to

any pair of concepts. The presented probabilistic approach is appropriate for analysis of data sets that contains uncertain relationships, such as links derived by text mining, since the confidence in the prediction can be plugged into the reliability measure. It should be noted that the key contribution of the paper is in testing the applicability of the network reliability measure on real datasets.

Tong *et al*. (2007b) have generalized CEPS to work with attributed graphs where vertices are labeled by a type from a set of vertex types (e.g., manager, accountant, etc.). Their goal is to find subgraphs that exactly match or nearly match a user query pattern and present them to the user in their "goodness" order. Their algorithm was used by Chau *et al*. (2008). Chau *et al*. (2008) developed the Graphite system that enables users to construct a query pattern, find all the exact and near matches and visualize the results. Tong *et al*. (2008b) also tracked the node centrality and proximity on a time-evolving bipartite graph (Kozen, 1992). They worked on a time-evolving author-conference network. Their work has two goals: to monitor the centrality of a node and to monitor the proximity of two nodes or sets of nodes.

Most of the existing proximity measurements only take into consideration the link structure of the graph and pay no attention to any side information such as feedback information. For example, the existing proximity measurements can answer a question like: What are the most similar conferences to "conference x"? In case a user dislikes "conference y" then by incorporating such side information, the question will be: What are the most similar conferences to "conference x" but not similar to "conference y". Tong *et al*. (2008c) proposed a method that incorporates the like/dislike side information when measuring proximity on graphs. The method uses the side information to bias the structure of the graph and is based on RWR. The authors proposed a fast solution for unipartite graphs. The proposed method is well-suited in many applications including CEPS. In CEPS, the original RWR can be replaced by the proposed method in order to deal with the side information. Tong *et al*. (2009) also proposed a method that incorporates side information but for bipartite graphs.

## CONCLUSION

This study presented a literature review on the connection subgraph problem. Connection subgraphs are useful in many applications such as ranking search results, discovering connections between criminals or terrorists, identifying connections between two genes, exploiting relationships to sell products and visualizing large graphs. One future research direction is developing an algorithm that finds connection subgraphs between n nodes by considering graphs with many types of edges or graphs with many types of vertices and edges. A second research direction is using parallelism to retrieve connection subgraphs faster from large graphs. A third future research direction is finding methods to enhance the available graph visualization tools in all aspects such as summarization, graphic design and display. One interesting subarea is visualizing RDF graphs.

## ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at King Saud university for funding the work through the research group No. 'RGP-VPP-067'. Also, the Authors wish to thank the Research Center - College of Computer and Information Sciences - King Saud University for their support of this work.

## REFERENCES

Aggarwal, C.C. and H. Wang, 2010. Managing and Mining Graph Data. Springer, New York, USA., ISBN-13: 9781441960443, pp: 610.

Ahmad, B. and H.K. Rizwan, 2005. Protein folding: From hypothesis driven to data mining. Pak. J. Biol. Sci., 8: 487-492.

Asur, S. and S. Parthasarathy, 2009. A viewpoint-based approach for interaction graph analysis. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDDM'09), New York, USA., pp: 79-88.

Bapat, R.B., 2010. Graphs and Matrices. Springer, New York, USA., ISBN-13: 9781848829800, pp: 171.

Berkhin, P., 2006. A Survey of Clustering Data Mining Techniques. In: Grouping Multidimensional Data, Nicholas, K. and Teboulle (Eds.). Springer, New Mexico, pp: 25-71.

Bose, P., W. Lenhart and G. Liotta, 1996. Characterizing proximity trees. Algorithmica, 16: 83-110.

Chakrabarti, S., 2003. Mining the Web: Discovering Knowledge from Hypertext Data. 1st Ed., Morgan Kaufmann Publishers, San Francisco, USA., ISBN-13: 978-1558607545, pp: 344.

Chau, D.H., C. Faloutsos, H. Tong, J.I. Hong, B. Gallagher and T. Eliassi-Rad, 2008. GRAPHITE: A visual query system for large graphs. Proceedings of the IEEE International Conference on Data Mining Workshops, Dec. 15-19, Pisa, Italy, pp: 963-966.

Chen, E., X. Chen, C.Y. Sheu and T. Qian, 2006. An evolutionary computational method for N-connection subgraph discovery. Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06), Arlington, VA., pp: 169-178.

Cheng, J., Y. Ke, W. Ng and J.X. Yu, 2009. Context-aware object connection discovery in large graphs. Proceedings of the 25th International Conference on Data Engineering, March 29-April 2, Shanghai, China, pp: 856-867.

Conrad, J., C.P. Gomes, W.J.V. Hoeve, A. Sabharwal and J. Suter, 2007. Connections in networks: Hardness of feasibility versus optimality. Proceedings of the 4th International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems, May 23-26, Brussels, Belgium, pp: 16-28.

Cook, D.J. and L.B. Holder, 2006. Mining Graph Data. John Wiley and Sons, Hoboken, New Jersey, ISBN-13: 978-0471731900, pp: 500.

Deo, N., 2004. Graph Theory with Applications to Engineering and Computer Science. Prentice Hall of India Pvt. Ltd., India, ISBN-13: 978-8120301450.

Eberhart, R. and Y. Shi, 2007. Computational Intelligence: Concepts to Implementations. 1st Ed., Morgan Kaufmann, Burlington, USA., ISBN-13: 978-1558607590, pp: 496.

Faloutsos, C., K.S. Mccurley and A. Tomkins, 2004a. Connection subgraphs in social networks. Proceedings of the Workshop on Link Analysis, Counterterrorism and Privacy, SIAM International Conference on Data Mining, March 15, 2004, California, USA., pp: 1-12.

Faloutsos, C., K.S. Mccurley and A. Tomkins, 2004b. Fast discovery of connection subgraphs. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDDM'04), ACM Press, New York USA., pp: 118-127.

Fournier, J.C., 2009. Graph Theory and Applications. ISTE Ltd. and John Wiley and Sons, USA., ISBN-13: 9781848210707, pp: 82.

Gabriel, K.R. and R.R. Sokal, 1969. A new statistical approach to geographical analysis. Syst. Zool., 18: 54-64.

Getoor, L., 2003. Link mining: A new data mining challenge. SIGKDD Explor. Newslett., 5: 84-89.

Getoor, L. and C.P. Diehl, 2005. Link mining: A survey. SIGKDD Explor. Newslett., 7: 3-12.

Hafez, A., J.S. Deogun and V.V. Raghavan, 1999. The item-set tree: A data structure for data mining. Proceedings of the 1st International Conference on Data Warehousing and Knowledge Discovery (DaWaK'99), Springer-Verlag, pp: 183-192.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publishers, San Francisco, CA., ISBN-13: 978-1558609013, pp: 800.

Herman, I., G. Melancon and M.S. Marshall, 2000. Graph visualization and navigation in information visualization: A survey. IEEE Trans. Visual. Comput. Graphics, 6: 24-43.

Hintsanen, P., 2007. The most reliable subgraph problem. Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'07), Springer-Verlag Berlin, Heidelberg, USA., pp: 471-478.

Hintsanen, P. and H. Toivonen, 2008. Finding reliable subgraphs from large probabilistic graphs. Data Min. Knowledge Discovery, 17: 3-23.

Jin, W., R.K. Srihari and X. Wu, 2007a. Mining concept associations for knowledge discovery through concept chain queries. Proceedings of the 11th Pacific-Asia International Conference on Knowledge Discovery and Data Mining, May 20-23, Osaka, Japan, pp: 555-562.

Jin, W., R.K. Srihari, H.H. Ho and X. Wu, 2007b. Improving knowledge discovery in document collections through combining text retrieval and link analysis techniques. Proceedings of the 7th IEEE International Conference on Data Mining, Oct. 28-31, Omaha, Nebraska, USA., pp: 193-202.

Kandel, A., H. Bunke and M. Last, 2007. Applied Graph Theory in Computer Vision and Pattern Recognition. 1st Edn., Springer, New York, ISBN-13: 978-3540680192, pp: 266.

Koren, Y., S.C. North and C. Volinsky, 2006. Measuring and extracting proximity in networks. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 20-23, Philadelphia, Pennsylvania, USA., pp: 245-255.

Koren, Y., S.C. North and C. Volinsky, 2007. Measuring and extracting proximity graphs in networks. ACM Trans. Knowledge Discovery, 55: 1-13.

Kozen, D.C., 1992. The Design and Analysis of Algorithms. 1st Edn., Springer-Verlag, New York, ISBN-13: 978-0387976877, pp: 320.

Kubat, M., A. Hafez, V.V. Raghavan, J. Lekkala and W.K. Chen, 2003. Itemset trees for targeted association mining. IEEE Trans. Knowledge Data Eng., 15: 1522-1534.

Maimon, O. and L. Rokach, 2010. The Data Mining and Knowledge Discovery Handbook. 2nd Edn., Springer, New York, ISBN-13: 9780387098227, pp:1285.

Mokeddem, D. and H. Belbachir, 2009. A survey of distributed classification based ensemble data mining methods. J. Applied Sci., 9: 3739-3745.

Noel, S., V. Raghavan and C. Chu, 2003. Document Clustering Visualization and Retrieval via Link Mining. In: Clustering and Information Retrieval, Wu, W., H. Xiong and S. Shekhar (Eds.)., Kluwer, Boston, MA, USA., pp: 161-194.

Phyu, T.N., 2009. Survey of classification techniques in data mining. Proceedings of the International MultiConference of Engineers and Computer Scientists, March 18-20, Hong Kong, pp: 727-731.

Raghavan, V. V. and K. Birchard, 1979. A clustering strategy based on a formalism of the reproductive process in natural systems. Proceedings of the 2nd Annual International ACM SIGIR Conference on Information Storage and Retrieval: Information Implications into the Eighties (ISRIIE'79), USA., pp: 10-22.

Raghavan, V.V. and C.T. Yu, 1981. A comparison of the stability characteristics of some graph theoretic clustering methods. Proceedings of the Conference on Pattern Analysis and Machine Intelligence, Jan. 27, IEEE, pp: 393-402.

Raghavan, V.V. and A. Hafez, 2000. Dynamic data mining. Proceedings of the 13th International Conference on Industrial and Engineering Application of Artificial Intelligence and Expert Systems, June 191-22, Springer Verlag, pp: 220-229.

Ramakrishnan, C., W.H. Milnor, M. Perry and A.P. Sheth, 2005. Discovering informative connection subgraphs in multi-relational graphs. SIGKDD Explor. Newslett., 7: 56-63.

Rodrigues, J.F., H. Tong, A.J.M. Traina, C. Faloutsos and J. Leskovec, 2006. GMine: A system for scalable, interactive graph visualization and mining. Proceedings of the 32nd International Conference on Very Large Databases (VLDB'06), VLDB Endowment Inc., USA., pp: 1195-1198.

Sevon, P., L. Eronen, P. Hintsanen, K. Kulovesi and H. Toivonen, 2006. Link discovery in graphs derived from biological databases. Proceedings of Data Integration in the Life Sciences, 3rd International Workshop, July 20-22, Hinxton, UK., pp: 35-49.

Sevon, P. and L. Eronen, 2008. Subgraph queries by context-free grammars. J. Integr. Bioinform., Vol. 5, No. 2. 10.2390/biecoll-jib-2008-100

Sharma, S., S. Khare and S. Sharma, 2007. Measuring the Interestingness of classification rules. Asian J. Inform. Manage., 1: 43-49.

Shehata, S., F. Karray and M. Kamel, 2006. Enhancing text retrieval performance using conceptual ontological graph. Proceedings of the 6th IEEE International Conference on Data Mining Workshops, Dec. 18-22, Hong Kong, China, pp: 39-44.

Srihari, R.K., S. Lamkhede and A. Bhasin, 2005a. Unapparent information revelation: A concept chain graph approach. Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05), ACM Press, USA., pp: 329-330.

Srihari, R.K., S. Lamkhede, A. Bhasin and W. Dai, 2005b. Contextual information retrieval using concept chain graphs. Proceedings of the Workshop on Context-based Information Retrieval, July 5, Paris, France, pp: 8-12.

Srihari, R.K., L. Xu and T. Saxena, 2007. Use of ranked cross document evidence trails for hypothesis generation. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 12-15, ACM Press, San Jose, California, USA., pp: 677-686.

Thakur, R.S., R.C. Jain and K.R. Pardasani, 2007. Fast algorithm for mining multi-level association rules in large databases. Asian J. Inform. Manage., 1: 19-26.

Tiwari, A., R.K. Gupta and D.P. Agrawal, 2010. A survey on frequent pattern mining: Current status and challenging issues. Inform. Technol. J., 9: 1278-1293.

Tong, H. and C. Faloutsos, 2006. Center-piece subgraphs: Problem definition and fast solutions. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 20-23, ACM Press, Philadelphia, Pennsylvania, USA., pp: 404-413.

Tong, H., C. Faloutsos and J.Y. Pan, 2006. Fast random walk with restart and its applications. Proceedings of the 6th International Conference on Data Mining, Dec. 8, IEEE Computer Society Washington, DC., USA., pp: 613-622.

Tong, H., Y. Koren and C. Faloutsos, 2007a. Fast direction-aware proximity for graph mining. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 12-15, San Jose, California, USA., pp: 747-756.

Tong, H., C. Faloutsos, B. Gallagher and T. Eliassi-Rad, 2007b. Fast best-effort pattern matching in large attributed graphs. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 12-15, San Jose, California, USA., pp: 737-746.

Tong, H., C. Faloutsos and J.Y. Pan, 2008a. Random walk with restart: Fast solutions and applications. Knowledge Inform. Syst., 14: 327-346.

Tong, H., S. Papadimitriou, P.S. Yu and C. Faloutsos, 2008b. Proximity tracking on time-evolving bipartite graphs. Proceedings of the SIAM Conference on Data Mining, April 24-26, Hyatt Regency Hotel Atlanta, Georgia, pp: 704-715.

Tong, H., H. Qu and H. Jamjoom, 2008c. Measuring proximity on graphs with side information. Proceedings of the 8th IEEE International Conference on Data Mining, Dec. 15-19, Pisa, Italy, pp: 598-607.

Tong, H., H. Qu, H. Jamjoom and C. Faloutsos, 2009. iPoG: Fast interactive proximity querying on graphs. Proceedings of the 18th ACM Conference on Information and Knowledge Management, Nov. 2-6, Hong Kong, China, pp: 1673-1676.

Toussaint, G., 1980. The relative neighborhood graph of finite planar set. Pattern Recognit., 12: 261-268.

Vast, S., P. Dupont and Y. Deville, 2005. Automatic extraction of relevant nodes in biochemical networks. Proceedings of the Learning and Bioinformatics Workshop, 7th Conference Speaking on Machine Learning (CSML'05), Cape Town, Nice, pp: 21-31.

Velmurugan, T. and T. Santhanam, 2011. A survey of partition based clustering algorithms in data mining: An experimental approach. Inform. Technol. J., 10: 478-484.

W3 ORG, 2004. Resource Description Framework (RDF): Concepts and abstract syntax. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/#section-rdf-graph.

Wallis, W.D., 2007. A Beginner's Guide to Graph Theory. 2nd Edn., Birkhauser Boston, New York, USA., ISBN-13: 9780817644840, pp: 260.

Yoon, J.P. and V.V. Raghavan, 2000. Multi-level scheme extraction for heterogeneous semi-structured data. Proceedings of the 1st International Conference on Web-Age Information Management, June 21-23, China, pp: 411-422.

Yu, P.S., J. Han and C. Faloutsos, 2010. Link Mining: Models, Algorithms and Applications. 1st Edn., Springer, New York, USA., ISBN-13: 978-1441965141, pp: 430.