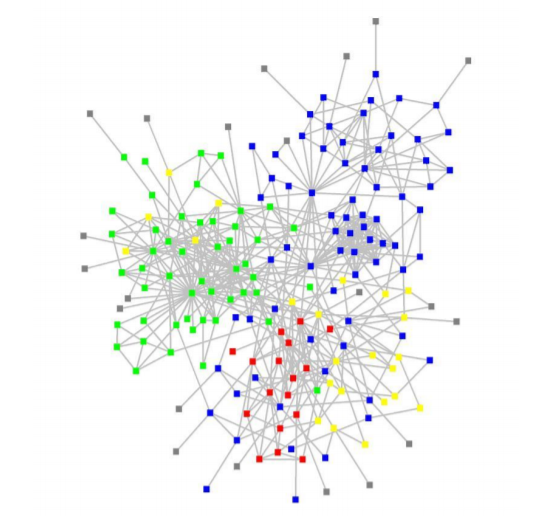
**LINK PREDICTION IN SOCIAL NETWORKS**

**Abstract:**

The zoom of social networks shows the increasing popularity of those services among the users. the expansion of social networks occurs as a results of adding new users and new links between users. The emergence of latest links has primacy within the study of social network evolution. Thus, predicting /recommending future links in social networks has attracted an excellent deal of attention. Link prediction has many applications and, it offers many benefits to the users of social networking services like providing fast and accurate recommendations or suggestions to the users. However, highly structured massive real-world networks involving heterogeneous entities with complex associations have added new challenges to link prediction research thanks to various factors like sparsity, complexity, size, time-dependent nature of the networks.

**Introduction :**

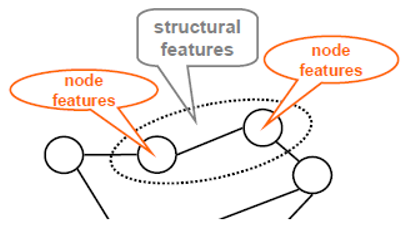
A social organization consists of nodes(Individuals or Organizations) and nodes are connected by differing types of relationships. a collection of social actors or nodes(such as individuals or organizations) and a collection of the dyadic ties between these nodes constitute a social network. as an example scientists in a very discipline, employees in a very large company, business leaders may be thought as nodes in a very network and coauthors of a paper, acting on a project, serve together on board may be thought as edges respectively. the thought behind Social Networks is to form opportunities to develop friendships, share information and promote business in a very network. OSN  
like Facebook and Twitter became important a part of way of life of countless  
people. the big growth and dynamics of those networks has led to many  
researches that examine the network properties i.e. structural and behavioral  
properties of huge scale social networks.



**Link prediction problem :**

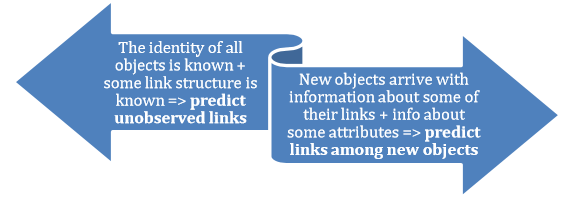
Link prediction problem is equal to network structure prediction problem, especially if we are considering social networks. There are several types of well-know link prediction methods that based on:

* node information
* structural information.



So, in social network relations among data are represented as a graph structure, where each node represents an information and a link represents a relation between two data, in other words nodes represent constituent elements and links represent relations among them. additionally, each node can even have an associated vector-structured data within the network model. compared to machine learning standard tasks settings: data is represented as tables, where rows represent observations and columns represent features/attributes.

Moreover, let’s distinguish two link prediction tasks in networked data:



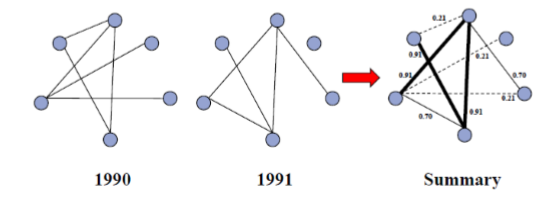
**Related Studies :**

Classification based on features of entities: entity attributes and relational (graph-based) features (indirect relations).

Attributes: number of neighbors, interests, topic model, affiliations, demographic data (geographical location).

Graph-Based Features: length of shortest path, neighborhood overlap, relative importance, mean first passage time.

Directed Graphical Models vs. Undirected Graphical Models

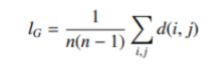


Application of probabilistic models to link prediction task

Graph Theory :

Graph Theory or Network Theory is a mathematical approach used in a way to foresee future possibilities of new link establishments in a current network. It consists of a combination of various mathematical concepts such as shortest paths in a network which states the approach of the nodes in a network such that shorter the distance between multiple nodes, the more likely that the nodes may connect and form a network in future. Certain multi-relational social networks like facebook take profile viewing, existing friend circles and known people network into consideration. This approach helps more in providing future friend recommendations to a user in a network.

Closeness centrality is another concept of Graph Theory which is regarded as the measure of how much a node is close to other nodes in a network. It is different from the shortest paths concept in a way that Closeness centrality is defined as the average length of all the shortest paths between a particular node and other nodes in a network. While Shortest Paths may predict only formation of a small social network, Closeness centrality provides a clear image of the evolution of a small network into a large network and vice versa. Let d(i,j) denote the shortest path between nodes i and j in a graph having n nodes. If i is not equal to j then L is defined as:



Probabilistic Approaches :

In principle, probabilistic approaches attempt to estimate the likelihood of potential links. The potential links with higher probabilities are more likely to happen than the links with lower probabilities. In contrast, the models like exponential random graphs are wont to estimate probabilistic models for the entire network. Thus, the probabilistic approaches may be put into two groups which models estimate probabilities of individual potential links and models estimate probabilities of potential structures of a current network. Besides that, we've to notice that the probabilistic methods are mostly supported the graph theoretical approaches described within the proceeding section.

In recent years, there has been growing interest in exponential random graph models for social networks. The exponential random graph models are a well-liked approach to estimate probabilistic models for an entire network using global features of a network, nodes and edges. These models are built upon statistical models which permit to inference about whether certain network substructures, often represented within the model by one or a tiny low number of parameters, are more commonly observed in the network than could be expected accidentally. we are able to then develop hypotheses about the social processes which may produce these structural properties. Exponential random graph models is a shot to create plausible models for networks by overcoming limitations of early graph theoretic approaches. The general form of the exponential random graph model for an observed graph Y is:



Where,

(1) Pr(Y = y) is the probability of Y taking the form y,

(2) ⴄA is the parameter corresponding to the configuration A such as triangle,

connected triple, etc. (ⴄA is non-zero only if all pairs of variables in A are

assumed to be conditionally dependent given the rest of the graph),

(3) gA(y) = Q yi j2A Yi j is the network statistic corresponding to configuration

A; gA(y) = 1 if the configuration is observed in the network y, and is 0

otherwise. yi j is a random variable denote the existence of link between node

i and j,

(4) k is a normalizing quantity which ensures that is a proper probability

distribution. In general, exponential random graph models are a good solution

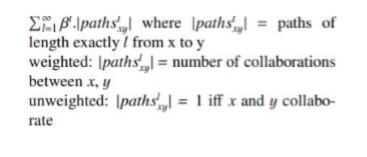
for study the evolution of small-world networks which have small number of

nodes and links.

For complex and large networks applying for large networks will have to pay high cost for computing.

Similarity Based Approaches :

Similarity based Approaches are used to show strong correlations between several measures of user activity and interchanging patterns. This influence the user centrality in a social network. The prominent characteristic of these Approaches are that they measure the similarity/dissimilarity of node pairs to assign a score or weight to them. More similar nodes get higher scores or weights and are more likely to link in the future. In general, the similarity based Approaches include neighbourhood methods such as Adam/Adar, Jaccard’s coefficient and Preferential Attachment and path-based methods such as Katz Index, Rooted PageRank and SimRank that are adapted from Graph Theory and are not suitable in case of link predictions in very large networks and are more suitable for small networks.



Katz beta index algorithm for computing collaborations between nodes x and y

One reason for its less suitability for large networks is because it relies solely on structural attributes of a network in its present form and not on supervised learning algorithms. Researchers propose a supervised learning framework that can be applied to existing algorithms such as Katz index and PageRank. In this method, the neighbourhood based and path based similarities of the nodes pairs are combinely termed as features. Then these features are represented by a vector and based on the past collaborations of the node pairs, the vector predicts if any future collaborations between the node pairs are possible or not.

**Machine Learning methods**

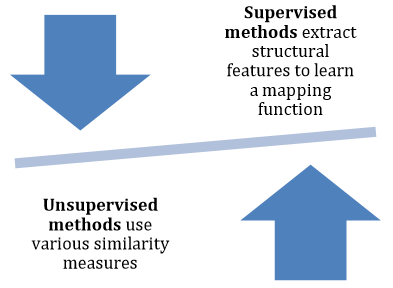
Machine Learning algorithms are termed as either “supervised” or “unsupervised”. Both supervised and unsupervised methods have been used in previous studies with different frameworks for Link Prediction. Even though both the methods have their pros and cons, the supervised methods have shown better performances compared to unsupervised methods. However, Machine Learning remains immense challenge due to the complexity and size of the networks as well as the temporal behaviours of the network.

Unsupervised learning methods :

Unsupervised learning methods are provided with unnamed classes in advance. These classes are then grouped based on the learning of the Unsupervised method in such a way that when it learns models from data that is presently unlabelled, it forms clusters of similar data and labels them as a class by giving it a suitable name. This method is used in the classification of instances in a way that shows the existence or non-existence of a link in a named class. If the link is present then there are higher chances of it inheriting the related properties of that class in the future.

Supervised learning methods :

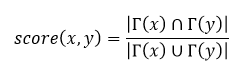
Supervised learning methods are one of the tasks carried out by Intelligent Systems. Thus, a large number of techniques have been carried out based on Artificial Intelligence (Logic-based techniques, Perceptron-based techniques) and Statistics (Bayesian Networks, Instance-based techniques). Unlike Unsupervised Learning Methods, Supervised learning methods have predetermined classes containing the labelled examples of link behaviour. These classes can be a finite set. The goal of the Supervised methods is to search and study the patterns formed by the node pairs and construct mathematical models based on it. These models are then evaluated on the basis of their predictive capacity in relation to the measure of the variance of the data itself. Available methods such as Decision Tree Induction, naïve Bayes, support vector machine and logistic regression are some examples of supervised learning techniques.

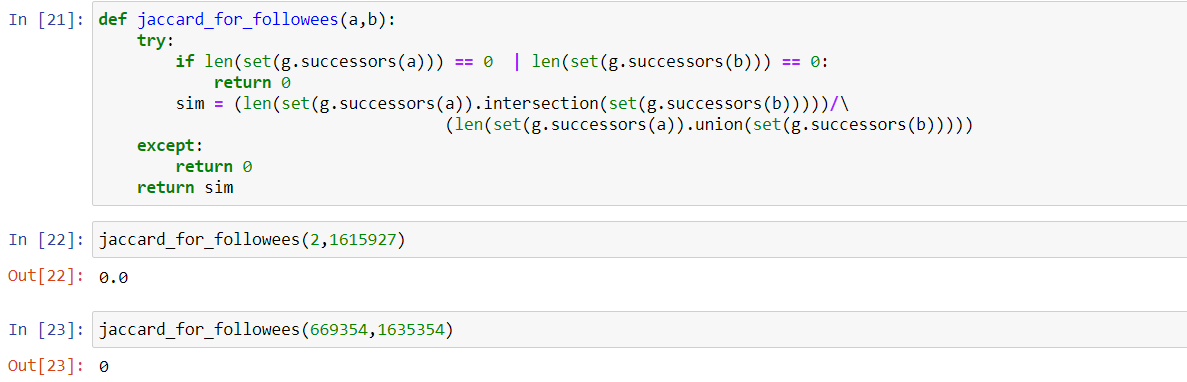


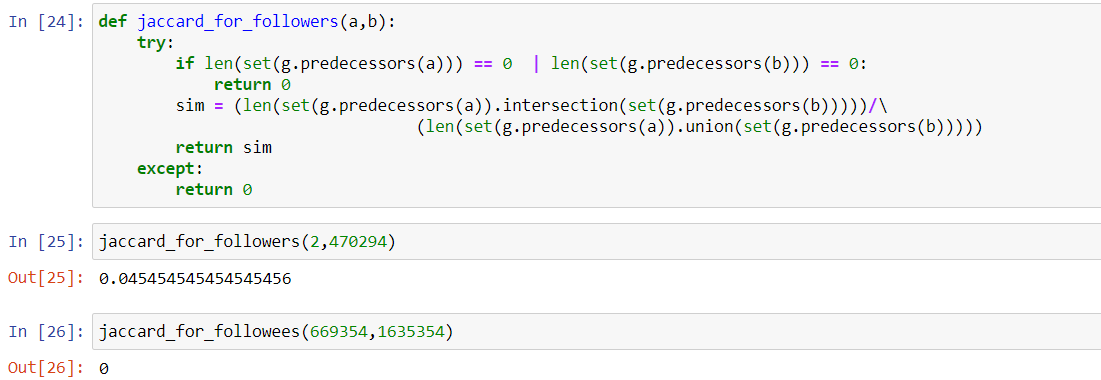
**Node Based Topological Patterns :**

Jaccard Distance:

The measure defines the issue that scientists may have in common neighbors because each has a lot of neighbors, but not because they are strongly related to each other:

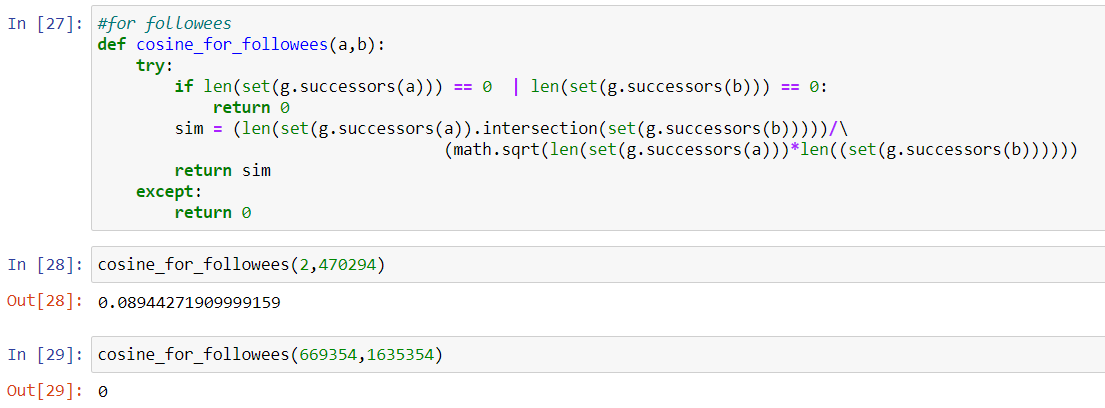
 (1)





Cosine distance

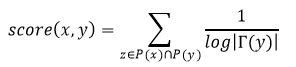
 (2)

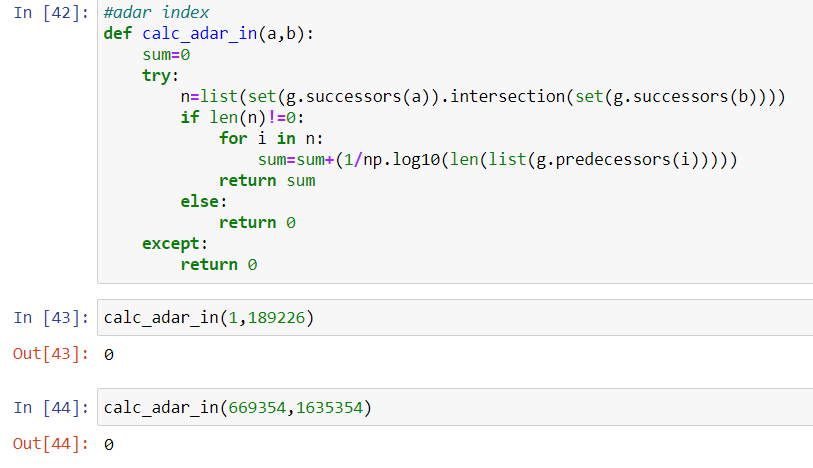




Adamic/Adar Index:

The measure allows counting common neighbors but gives more weight to neighbors that are not shared with many others:

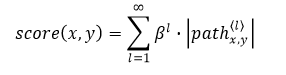
 (3)

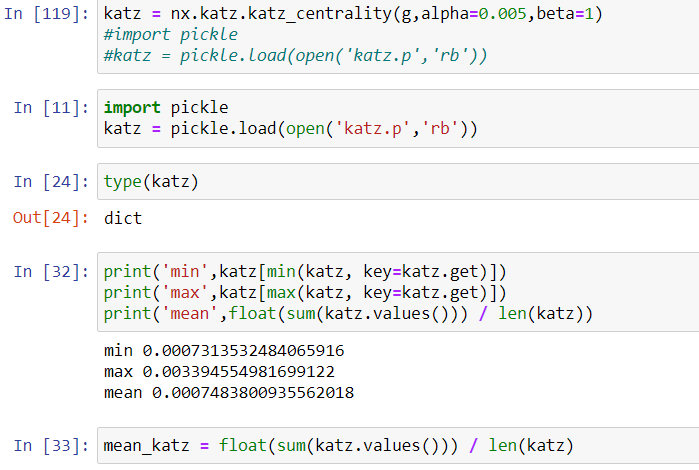


**Path Based Topological Patterns :**

Katz Centrality:

Measure that directly sums over this collection of paths, exponentially damped by length to count short paths more heavily. This notion leads to the measure where 𝑝𝑎𝑡ℎ𝑠𝑥,𝑦 𝑙 is the set of all length-l paths from 𝑥 to 𝑦, and 𝛽 > 0 is a parameter of the predictor. There are two variants of this Katz measure: (a) unweighted, in which 𝑝𝑎𝑡ℎ𝑠𝑥,𝑦 𝑙 = 1 and 𝑦 have collaborated and 0 otherwise, and (b) weighted, in which is the number of times that 𝑥 and 𝑦 have collaborated.





Hits Score :

A random walk on 𝐺 starts at a node 𝑥 and iteratively moves to a neighbor of 𝑥 chosen uniformly at random from the set 𝛤(𝑥). The hitting time 𝐻𝑥,𝑦 from 𝑥 to 𝑦 is the expected number of steps required for a random walk starting at x to reach y. Because the hitting time is not in general symmetric, it also is natural to consider the commute time 𝐶𝑥,𝑦 = 𝐻𝑥,𝑦 + 𝐻𝑦,𝑥. Both of these measures serve as natural proximity measures and hence (negated) can be used as 𝑠𝑐𝑜𝑟𝑒(𝑥,𝑦).



**Conclusion :**

Link prediction may be a relatively young research area and many open challenges remain. Further studies are required so as to grasp why some methods work better or worse than others looking on the network they're applied to. Studying which network structural properties result in better performance for every technique is an open research problem. additionally, only a few techniques adapt to the worldwide structure of the network and no technique adapts to the local structure of networks. the most difficulty when managing complex networks in practice is their size, which limits the styles of techniques which will be applied.

Link prediction remains an open research problem, given its importance in many applications. New techniques with better accuracy and performance trade-offs are expected to be proposed within the forthcoming future.

**References :**

1. <https://medium.com/@vgnshiyer/link-prediction-in-a-social-network-df230c3d85e6>
2. <https://ieeexplore.ieee.org/document/8985354/algorithms>
3. <https://ieeexplore.ieee.org/document/7058003>

**GitHub Link :**