Social Media Analytics

UNIT IV: LECTURE 01

Content

Overview of Social Media Analytics

Seven Layers of Social Media Analytics

Social Network Analysis

- Link Prediction
- Community Detection
- Influence Maximization

Text Analytics/Mining

- Text Categorization
- Document or Text Summarization
- Sentiment Analysis

Trend Analytics

Social Media Analytics: An Overview

Social Media Analytics: An Overview

- With the birth of social media, users through Internet can comment, express their opinions or send messages to each other at an incredible speed that was once thought as a dream by our ancestors.
- Some of the **popular social media applications** can be categorized as discussion forums, weblogs, social network sites, microblogs, wikis, image and/or video sharing platforms, and review sites.
- Studies reveal that nowadays users spent a major amount of their daily time using social media sites and the rate in this usage is **increasing year by year**.
- If we consider all the popular social media applications or platforms, the total number of users is significantly large and, in turn, the content generated through these users is immensely huge.

Social Media Analytics Process

- Social media analytics is one of the emerging topics in the area of research in data science and it is still in its infancy.
- Let us quickly explore the **social media analytics process** which comprises of **three stages** – data capturing, data understanding and data presentation.
- These stages are illustrated in Figure that shows the work done in each stage of the process.

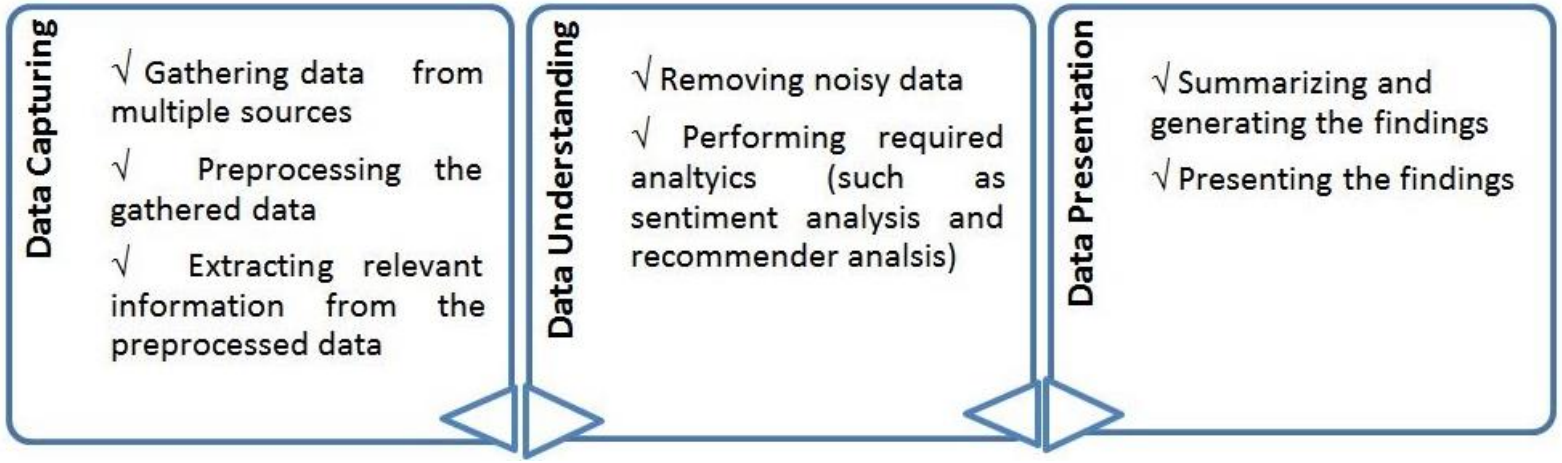


***Fig**: Social Media Analytics Process*

Social Media Analytics Process

1. Data Capturing

- Data capture, in simple terms, means **valid data identification**.
- In the data capturing stage a data analyst or a data scientist procures relevant social media data by "listening" to various sources of social media.
- The term listening here refers to the process of monitoring and assessing the social media content to understand the users' reactions, likes and dislikes, interests, reviews, and comments about a topic, agenda, brand, item or product.
- Capturing of data may comprise of **collecting data from various platforms** such as Facebook, Instagram, LinkedIn, Twitter, wikis, microblogs, and many such popular sites.
- Also, proper **preprocessing of data** is performed which is very crucial for proper analysis of data.
- **Extracting pertinent data** in this stage will allow more refined understanding of data in the next stage.

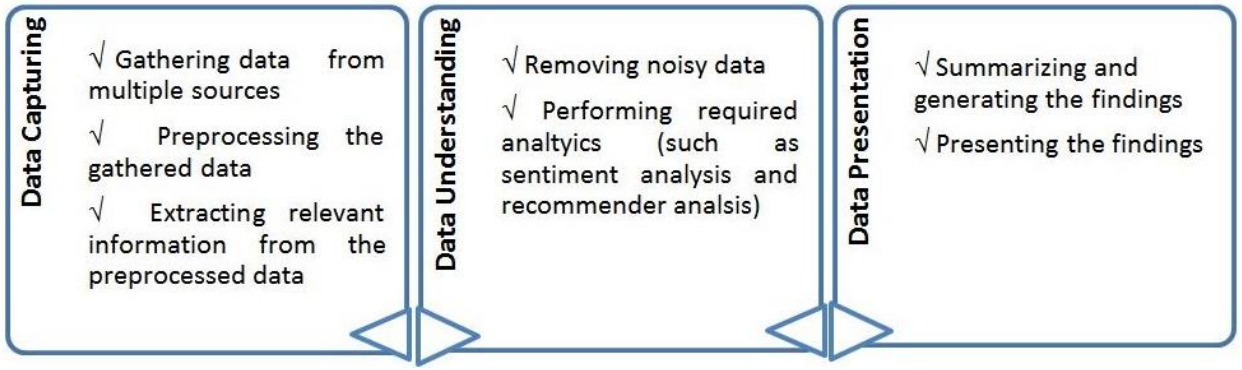


Fig: Social Media Analytics Process

Social Media Analytics Process

2. Data Understanding

- Data understanding is all about utilizing the data supplied from the data capturing stage and **analyzing the captured data for gaining meaningful insights**.
- Sometimes, before analyzing the data, the process of noise removal from data may be required for better accuracy in data analysis.
- Data analysis at this stage may involve several fields and techniques such as **statistical analysis, machine learning, deep learning, computer vision, natural language processing, and/or data mining**.

- The data understanding stage lies in the **middle** of the social media analytics process and forms the core and most important stage in this entire process.
- Once the analysis is performed, the analyzed data is further presented to the next stage, which is the data presentation stage.

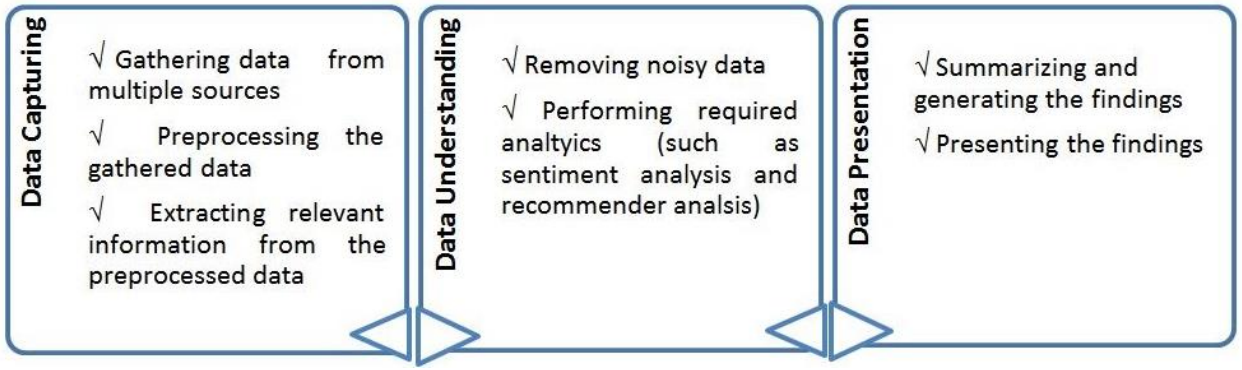


Fig: Social Media Analytics Process

Social Media Analytics Process

3. Data Presentation

- Data presentation is the final stage in the social media analytics process. In this stage, the **results are summarized and evaluated to gain significant insights**.
- These final results are then presented mostly using proper **data visualization tools** to present the output in an easy interpretable form.
- It is important to note that **data presentation is what the users get as output** at the end and hence, no matter how big the data is in volume, the data visualization graphic(s) should make the **output easily understandable** for the data analysts and data scientists.
- Interactive data visualization has led us to a new era of revolution where graphics have led us to **easy analysis and decision making.**
- The most challenging part however is to learn how data visualization works and which visualization tool serves the best purpose for analyzing precise information in a given case.

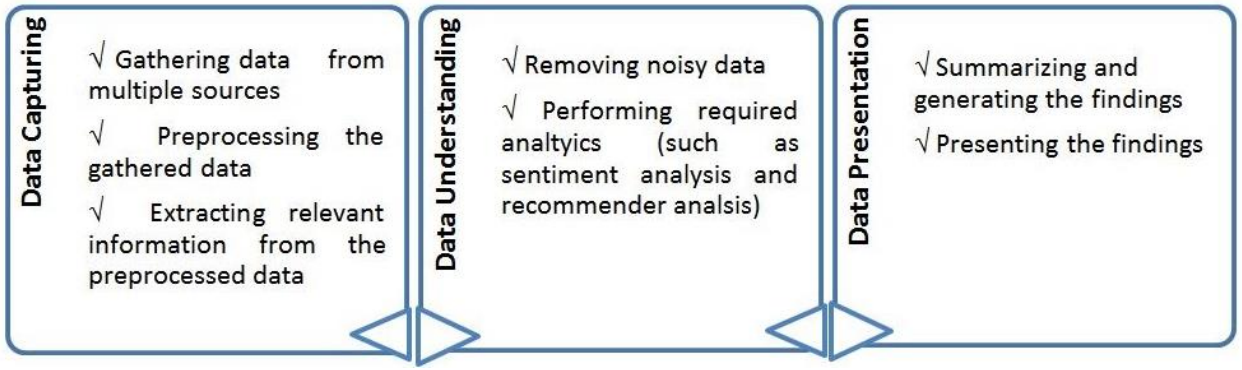


***Fig**: Social Media Analytics Process*

Seven Layers of Social Media Analytics

Seven layers of Social Media Analytics

- Social media is said to consist of **seven layers of data** that contain useful information that is often garnered for business intelligence.
- These seven layers of data may be either **visible** (say, textual data) or **invisible** (say, hyperlink network).
- The Figure beside shows the seven layers of data found in social media. All these seven layers play a vital role in contributing to social media input for gaining useful insights in businesses.

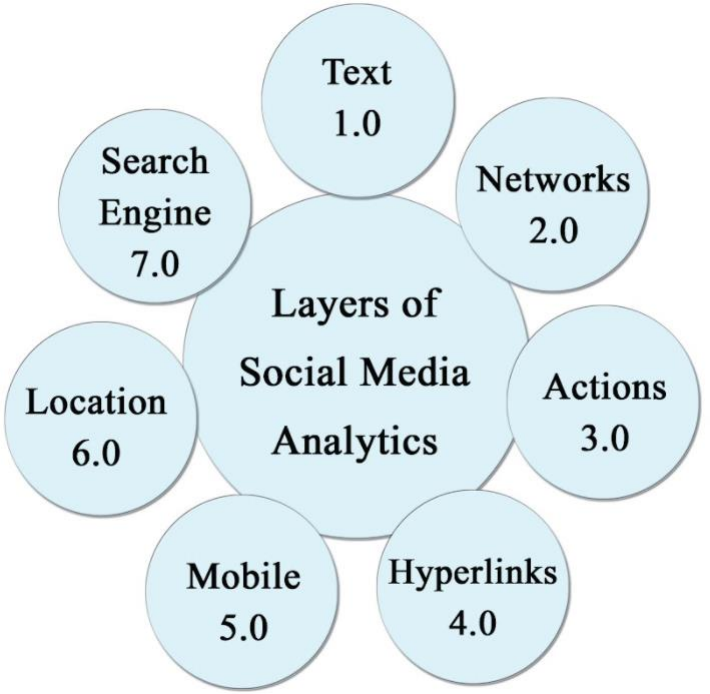


***Fig**: Seven Layers of Social Media Analytics Process*

Seven layers of Social Media Analytics

1. **Layer 1: Text** – The textual message of social media mainly include **tweets, textual posts, comments, status updates**, and **blog posts**. These texts are often in business analytics to identify user opinion or sentiment regarding particular product, topic, or individual.

2. **Layer 2: Networks** – Social media network analytics focuses on the **networking structure of the social media data** which indicates the connection between users based on the concept of friends and followers. Such connections are often found in various networking sites such as Facebook, Twitter, LinkedIn, and Instagram. The network analysis is mainly done using graph theory in which nodes are considered as the users and the edges are considered as the links or connection among users. Network analysis are often done to identify influential users, predict new links, and for various other analysis.

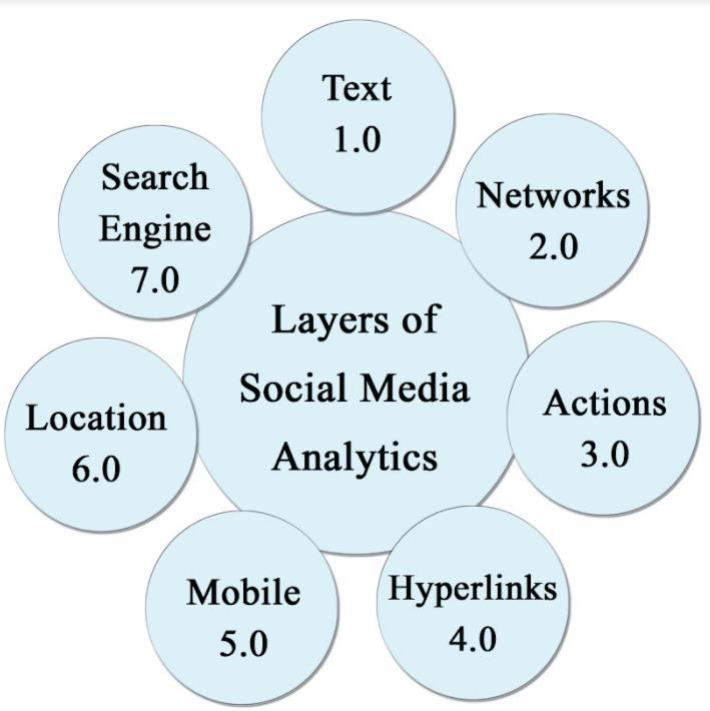


***Fig**: Seven Layers of Social Media Analytics Process*

Seven layers of Social Media Analytics

3. **Layer 3: Actions** – Social media actions mainly include the actions performed by users while using social media such as **clicking on like button, sharing posts, creating new events or groups, accepting a friend request**, and so on. Data analysts often carry out actions analytics using social media data for measuring various factors such as **popularity of a person or item, recent trends followed by users, and popularity of user groups.**

4. **Layer 4: Mobile** – **Mobile analytics** is comparatively a recent trend in social media analytics that focuses on analysis of user engagement with mobile applications. Mobile analytics are usually carried out for marketing analysis to attract those **users who are highly engaged with a mobile application**. Another common analysis carried out in mobile analytics is in-app analysis which concentrates on the kind of activities and interaction of users with an app.

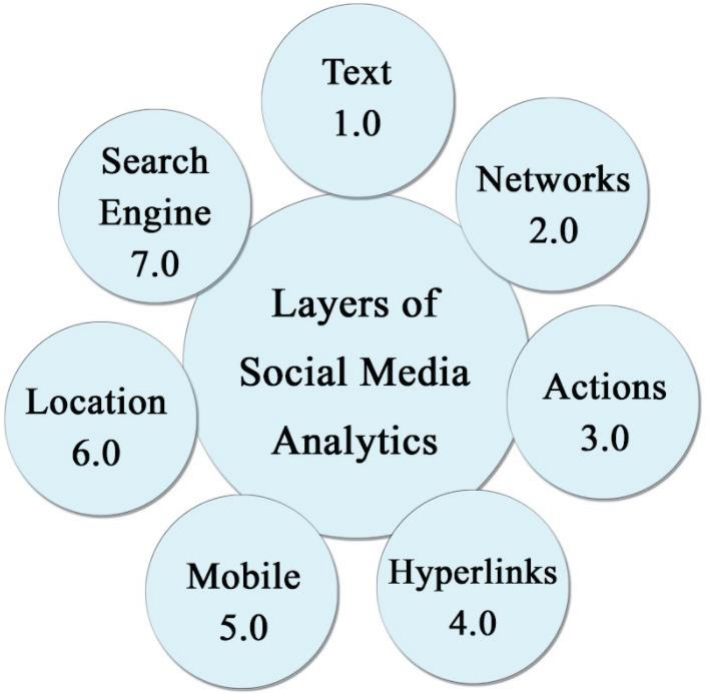


***Fig**: Seven Layers of Social Media Analytics Process*

Seven layers of Social Media Analytics

5. **Layer 5: Hyperlinks** - Hyperlinks are commonly found in almost all web pages that allow navigation of one web page to another. The hyperlink into a web page is called as **in-link** whereas the hyperlink out of a web-page is called as **out-links**. The number of in-links to a web page is referred to as **in-degree** whereas the number of out-links from a web page is referred to as **out-degree**. Mobile analytics is all about analyzing and interpreting social media hyperlinks.

6. **Layer 6: Location** - Location analytics is also known as **geospatial analysis** or simply **spatial analytics**. Location analytics is carried out to gain insight from the geographic content of social media data. Real-time location analytics is often carried out by data analysts for business intelligence. For instance, the **courier services** used by social media sites need to keep track of the locations of delivery in real-time.

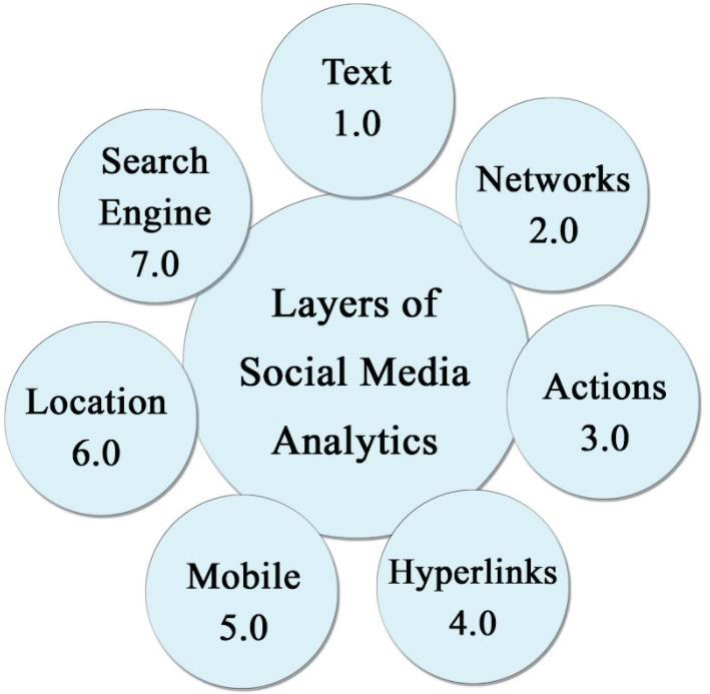


***Fig**: Seven Layers of Social Media Analytics Process*

Seven layers of Social Media Analytics

7. **Layer 7: Search engines** - Search engine analytics pays attention to analyzing historical search data to generate informative search engine statistics. These statistical results can then be used for **search engine optimization (SEO)** and **search engine marketing (SEM)**. A few of the tasks that involve search engine analytics include advertisement spending statistics, keyword monitoring, and trends analysis.

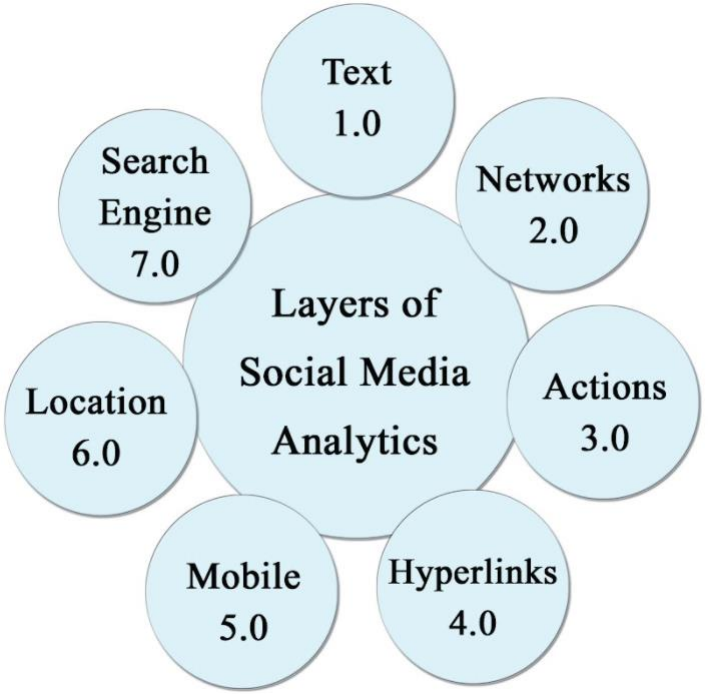


***Fig**: Seven Layers of Social Media Analytics Process*

Key Social Media Analytics

Key Social Media Analytics

In the context of social media analytics, there are mainly three methods of analysis that have a variety of applications.

These **three primary methods** for social media analytics mainly include:

a. **Social network analysis**

b. **Text analysis/mining, and,**

c. **Trend analytics**

Social Network Analysis

Social Network Analysis

- SNA emphasizes analyzing the **users** in a network (often referred to as **nodes**) and their **connections** among each other (often termed as **edges**).
- By concentrating on the **structure of connections** among users, SNA can help identify opinion leaders, influential users, or user communities in social media.
- Few of the mining issues dealt in social network analysis and mining include:
  - Link prediction
  - Community detection
  - Influence maximization
  - Expert finding, and,
  - Prediction of trust and distrust among individuals

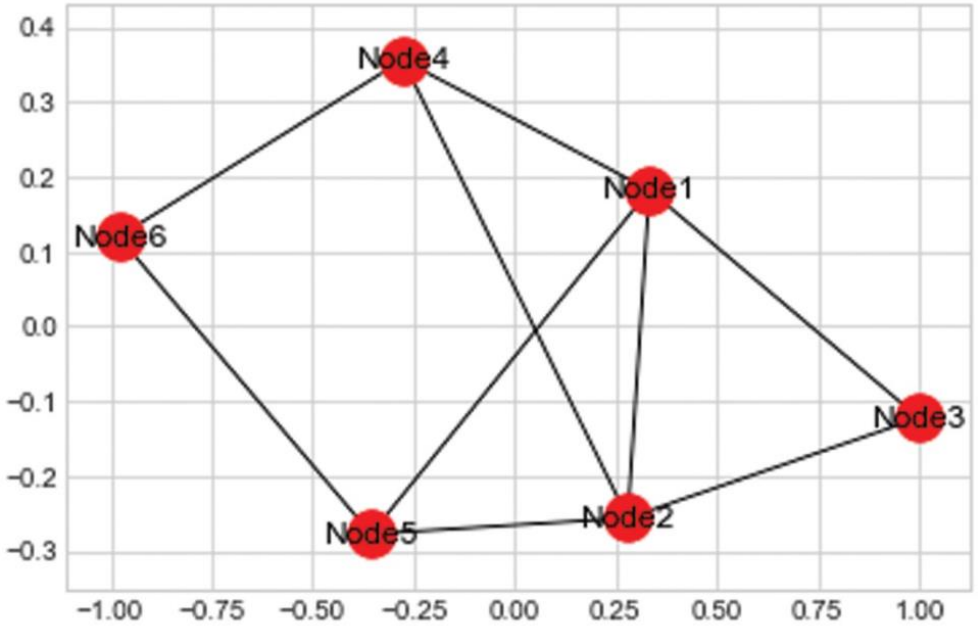
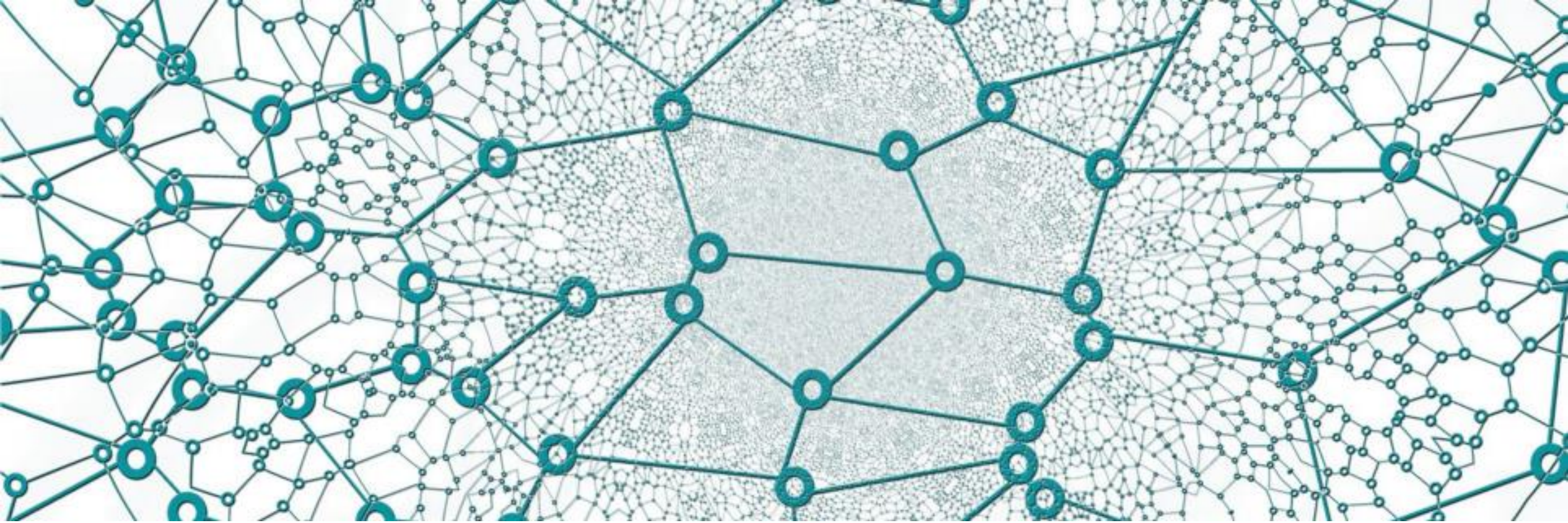


***Fig**: Users and their Connections in a Network*



Social Network Analysis:
Link Prediction

Social Network Analysis: Link Prediction

- One common research issue in social network analysis and mining is the link prediction problem which studies a **static snapshot** of the nodes and edges of a social network at a given time T1 and based on the study, predicts the future links of the social network for a future time T2.
- The link prediction problem is a common feature found in many social networking sites for possible **friends' suggestions** as found on Facebook or LinkedIn. This feature, in turn, allows a user to increase the personal or professional friends circle to broaden the social links and connections.

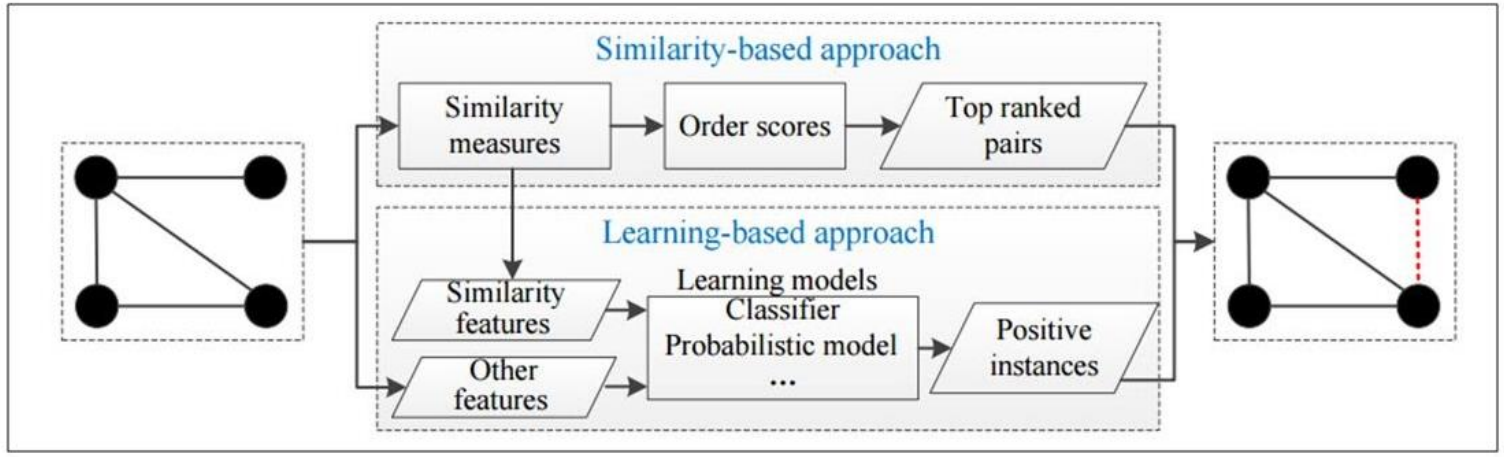


***Fig**: Link Prediction in Social Network*

Social Network Analysis: Link Prediction

- The Figure illustrates a standard link prediction framework that feeds a static social network as input and then applies either a **similarity-based approach** or a **learning-based approach** for prediction of future links in the social network.
- **Similarity-based approach of link prediction**: Calculates the similarities of non-connected pair of nodes in a social network and a score is accordingly assigned for each non-connected pair. Based on the descending order of similarity score, a list is prepared to choose the **top-N ranked links** from the list for link prediction.
- **Learning-based approach of link prediction:**
  - A classifier that uses some standard machine learning models to **assign a label that is binary** – positive or negative.
  - A positive value indicates that there is a chance of better connectivity between the nonconnected pair of nodes whereas a negative value indicates that there is very little chance of connectivity between the non-connected pair nodes.

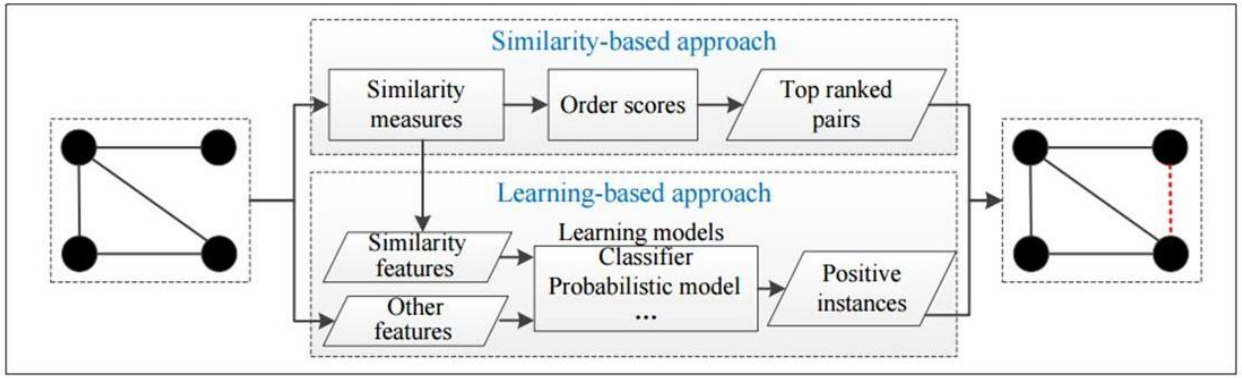


Fig: Link Prediction in Social Network

The Link Prediction Problem

- Given in the Figure is a 7 node graph and the unconnected node-pairs are AF, BD, BE, BG, and EG
- **Objective**: To predict whether there would be a link between any 2 unconnected nodes.
- Link prediction has a ton of use in real-world applications. Here are some of the important use cases of link prediction:
  - Predict which customers are likely to buy what products on online market places like Amazon.
  - Suggest interactions or collaborations between employees in an organization
  - Extract vital insights from terrorist networks

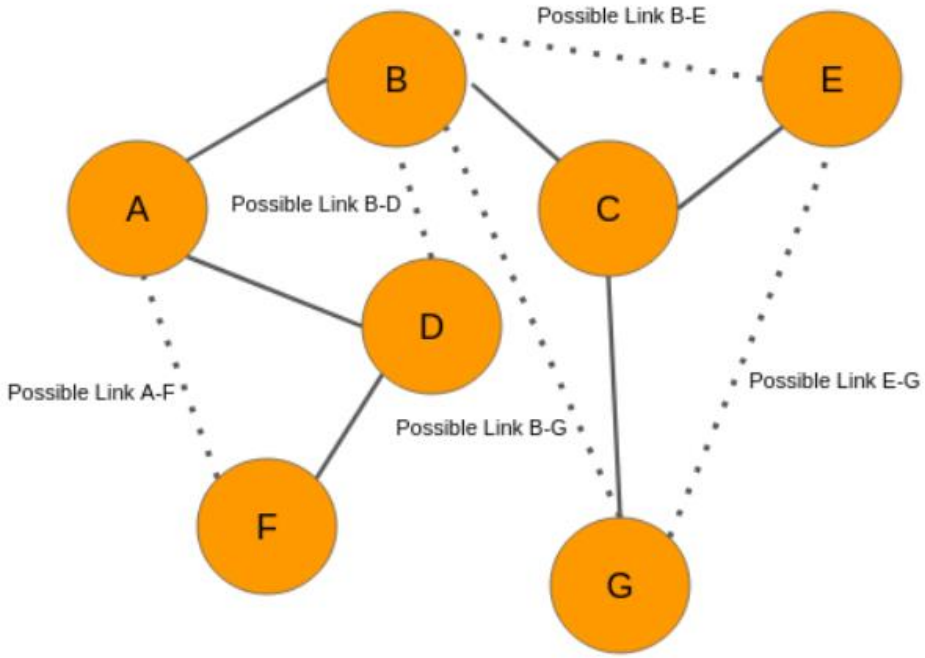


***Fig**: Social Network Graph*

Standard Link Prediction Techniques

- **Common Neighbors Method:**
  - A common approach to link prediction is to compute the number of common neighbors.
  - Entities with more neighbors in common are more likely to have a link. For nodes A and B, it can be computed as follows:

$$\text{Score}(A, B) = | \text{Neighbors}(A) \cap \text{Neighbors}(B) |$$

  - The approach used in common neighbors is very straightforward and easy to analyze.
  - Also, this approach is very effective as it has been experimentally evaluated many a times that this technique outperforms several other complex techniques used for link prediction.

Standard Link Prediction Techniques

- **Jaccard Coefficient Method:**
  - Jaccard's coefficient is another basic technique often used for prediction of links.
  - In case of Jaccard's coefficient, the probability of two nodes A and B to be connected in the near future is based on the score value given below:

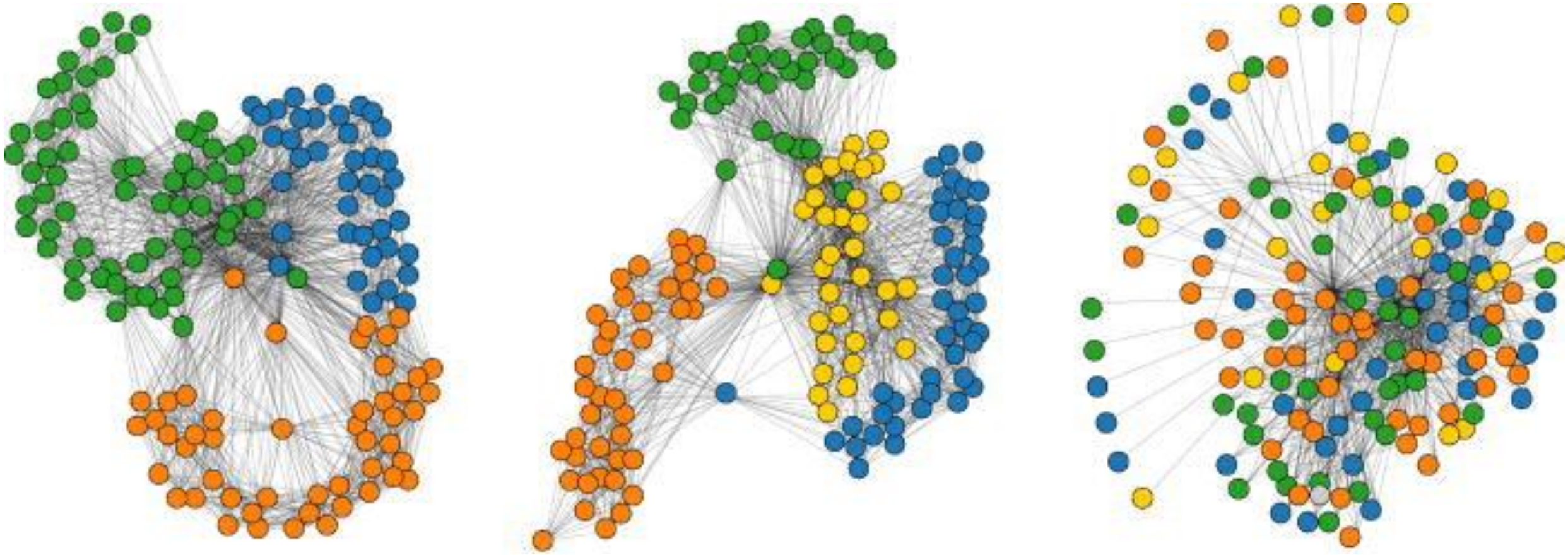
$$\text{Score}(A, B) = \frac{|\ \text{Neighbors}(A) \cap \text{Neighbors}(B)\ |}{|\ \text{Neighbors}(A) \cup \text{Neighbors}(B)\ |}$$

  - Here, in case of Jaccard's coefficient, the score between two non-connected nodes is calculated by considering the number of common neighbors (numerator) and simply divide this value by the total number of neighbors (denominator).

Standard Link Prediction Techniques

- Some other standard methods of link prediction include the following:
  - The *Adamic / Adar* approach
  - The *Preferential Attachment* approach
  - The *FriendLink* approach
  - The *LinkGyp* approach[1]

[1]***Reference***: Nandi, Gypsy, and Anjan Das. "An Efficient Link Prediction Technique in Social Networks based on Node Neighborhoods." International Journal of Advance Computer Science and Applications 9.6 (2018): 257-266.



Social Network Analysis: Community Detection

Social Network Analysis: Community Detection

- For community detection in social networks, a study is carried out to find the ***correlation between nodes in the network*** to assess the strength of the connection between nodes.
- With this, **intra-communities** are formed in which a group of users (or nodes) with close correlation belong to the same community and in **inter-communities**, users (or nodes) belong to different communities.
- A user belonging to the same community is expected to share **similar tastes, likes and dislikes**. This helps in the prediction of what products a user is likely to buy, which movie a user is likely to watch, what services a user may be interested in, and so on.

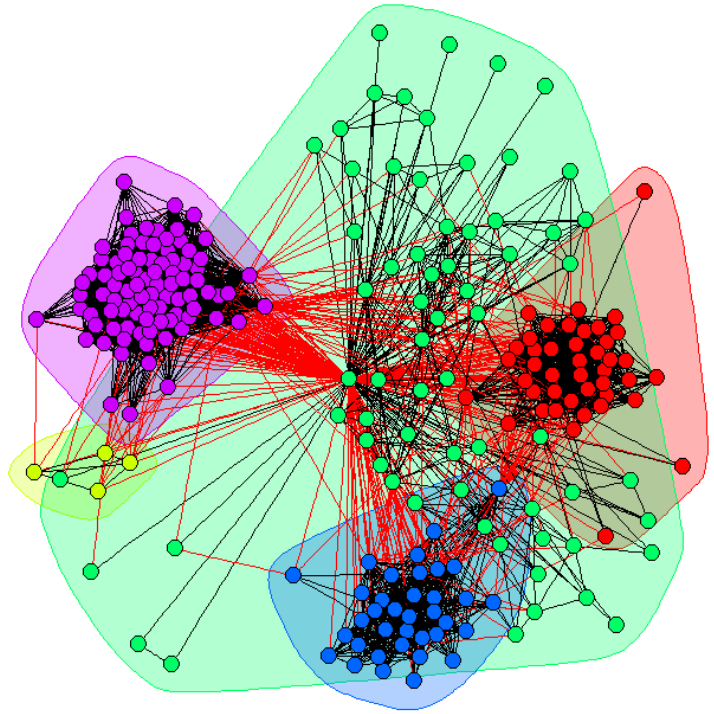


***Fig**: Community Detection in Social Networks*

Social Network Analysis: Community Detection

- Figure below shows the various categories of approaches followed for community detection in social networks.

- These **categories** are broadly divided into:
  - **The Traditional Clustering Community Detection methods**
  - **The Link-based Community Detection methods**
  - **The Topic-based Community Detection methods, and,**
  - **The Topic-link based Community Detection methods**

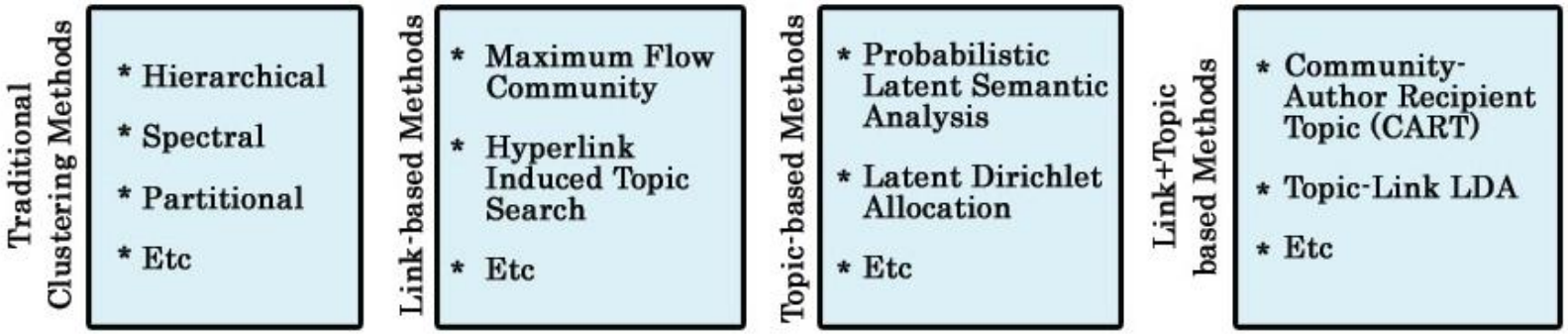
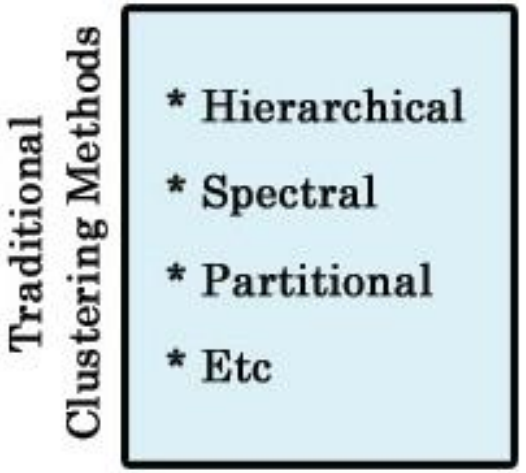


***Fig**: Various Categories of Community Detection Methods*

Community Detection: Various Categories

1. **Traditional Clustering Methods:**

- The various traditional clustering methods of community detection is mainly divided into hierarchical, spectral, and partitional methods.
- The **hierarchical clustering** method either gradually **merges or splits** the groups to create nested clusters.
- **Spectral clustering** creates groups of communities by using the spectral properties of the **similarity matrix**.
- **Partitioning clustering** divides all nodes into **n clusters**, where the value of n is provided as a parameter well in advance.



Community Detection: Various Categories

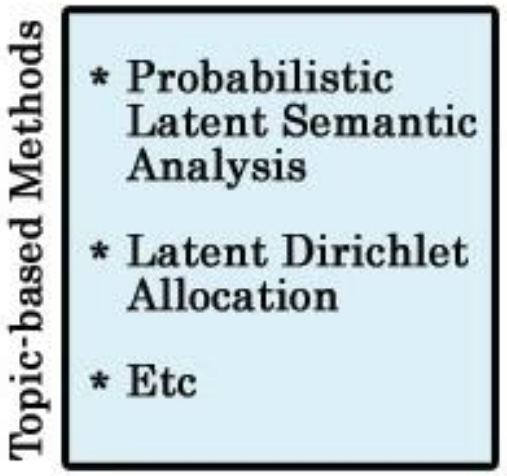
2. **Link-based Clustering Methods:**

- These community detection methods emphasize the **study of edges of the social network** in order to form communities.
- In link-based community detection methods, what is mainly explored is the **strength of connections between nodes** and not the basic semantics such as the common topic of interests or likings among nodes.
- Two standard link-based community detection methods (as shown in Figure) are **Hyperlink Induced Topic Search (HITS)** and **Maximum Flow Community (MFC).**

Community Detection: Various Categories

3. **Topic-based Clustering Methods:**

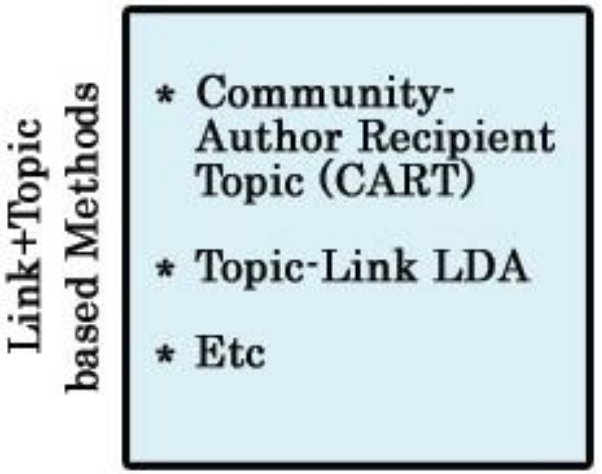
- These community detection methods emphasize the generation of communities based on the **common topic of interests**.
- In topic-based community detection methods, what is mainly explored is finding **communities that are topically similar**, and do not consider any emphasis on the strength of connections between nodes.
- Two standard topic-based community detection methods (as shown in Figure) are **Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet allocation (LDA).**

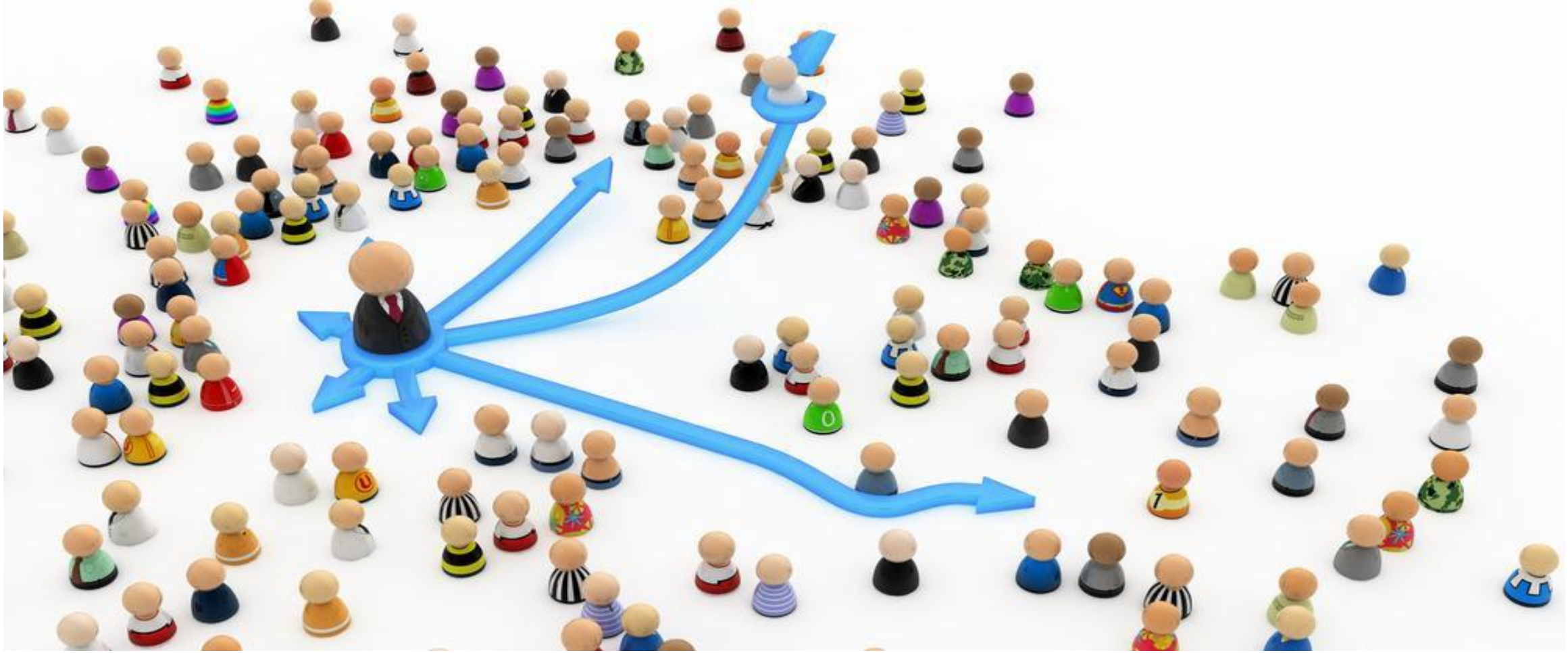


Community Detection: Various Categories

4. **Topic-Link based Clustering Methods:**

- These community detection methods are the **most common approaches used nowadays** for community detection in social networks.
- It is considered as a hybrid method as it considers both the strength of connections between nodes as well as finding communities that are topically similar.
- This method considers the disadvantages of using only one single method – link-based or topic-based, for community detection, and combine both the methods to give more accurate results.
- Two standard topic-link based community detection methods (as shown in Figure) are **Community-Author-Recipient-Topic (CART)** and **Topic-Link LDA.**





Social Network Analysis: Influence Maximization

Social Network Analysis: Influence Maximization

- Influence propagation is the task of choosing a set of proficient users who can prove to be very efficient for **viral marketing**.
- This set of efficient users in a social network is called a **seed set** which is considered as valuable nodes to target for promotion or publicity as these online users have the highest reach of spreading information.
- Indirectly, the seed set of users can help other users to decide in choosing as to which movie to watch, which political party to follow, which product to buy, which community to join, and so on.

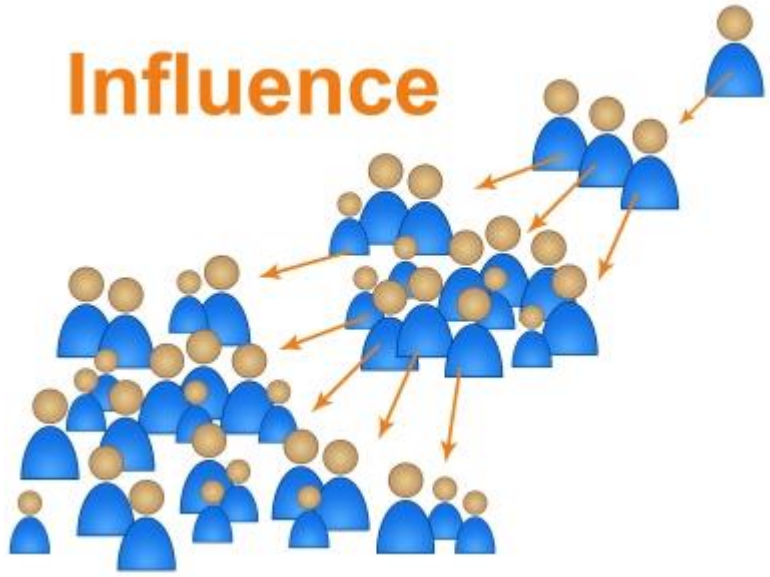


***Fig**: Influence Maximization in Social Networks*

Social Network Analysis: Influence Maximization

- Viral marketing is an effective tool being adapted by all companies and organizations for the publicity and promotion of their brands.
- For this, initially an influence maximization technique is used to find a set of few influential users of a social network, and influence those users about the goodness and usefulness of a product so that it can create a cascade of influence for buying the same product by the users' friends.
- The users' friends will again, in turn, recommend or publicize the same product to their friends, and this helps in easy **product promotion.**

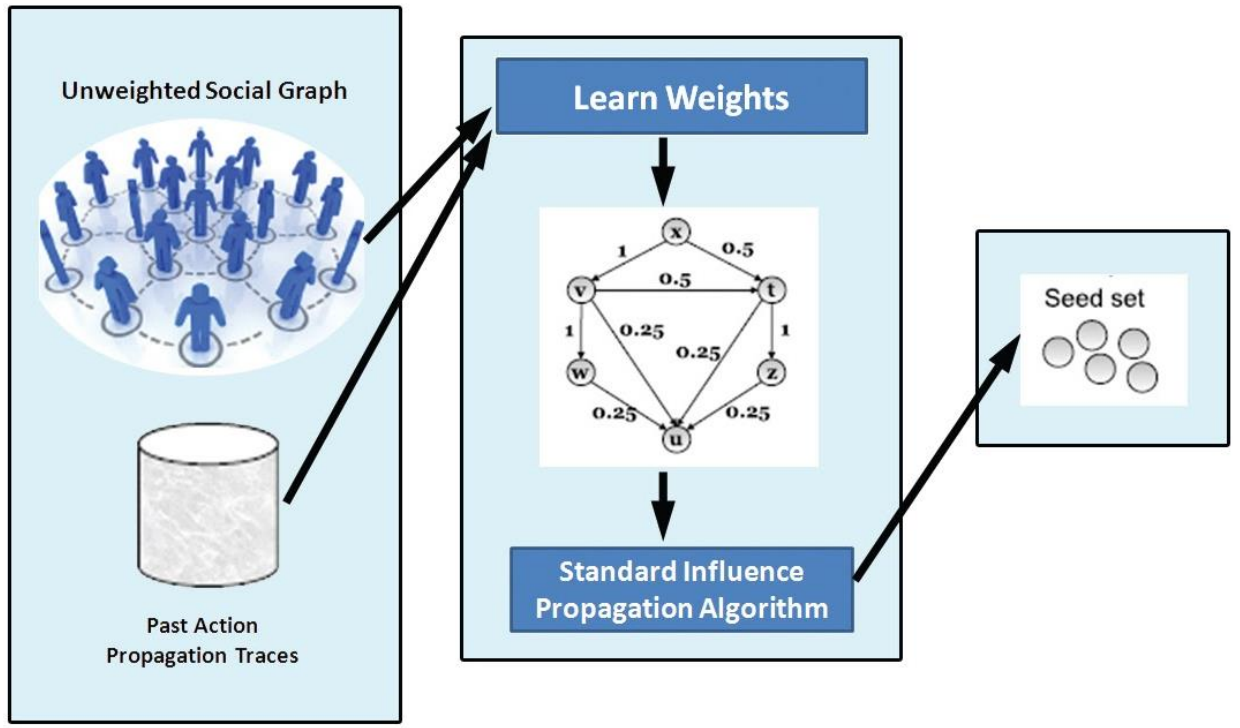


***Fig**: The Generic Influence Maximization Framework*

Social Network Analysis: Influence Maximization

- The Figure demonstrates the generic influence maximization model in which an unweighted social graph is fed as input to the model.
- The social graph contains **past action propagation traces** which are then used by an influence diffusion technique to learn the weights of each edge.
- Now the unweighted social graph is converted to a weighted graph which is again provided to a standard influence maximization algorithm to **generate the seed set** which is considered as output for the entire influence maximization model.

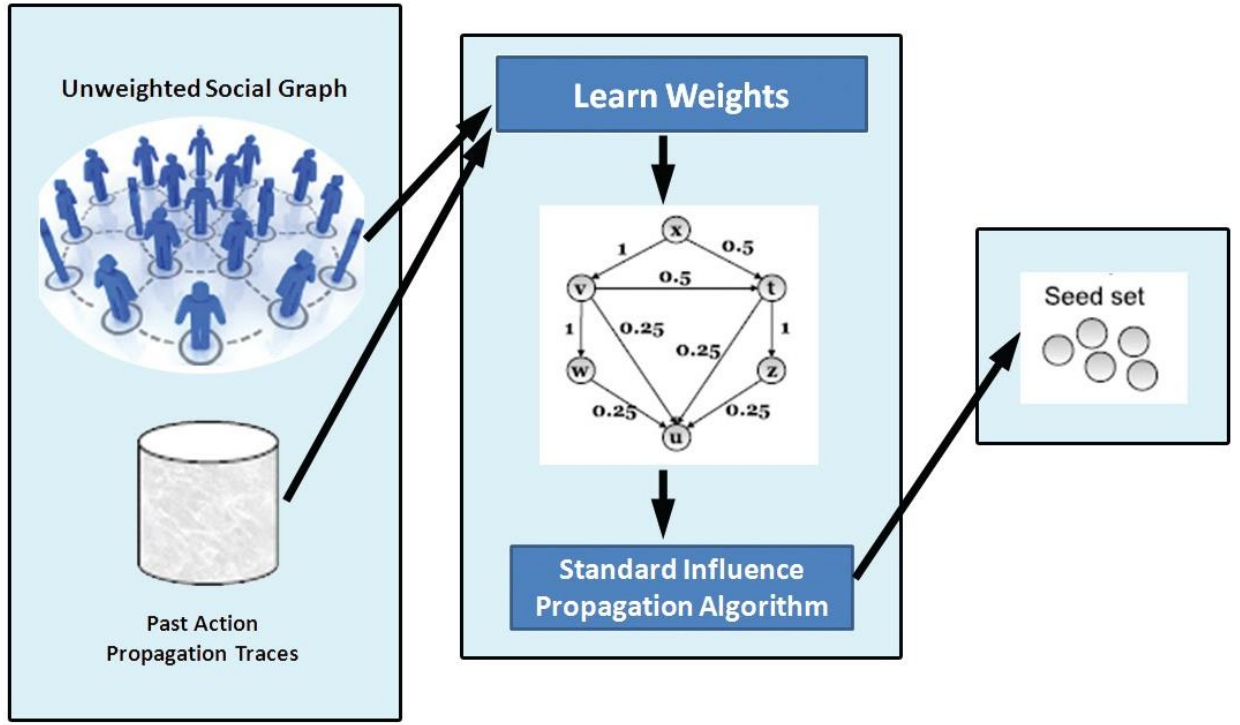


Fig: The Generic Influence Maximization Framework

Common Influence Maximization Technique

- **The High-Degree Centrality Approach:**
- Degree Centrality (DC) is based on the basic concept that nodes having higher direct connections with other nodes are considered to have higher degree of influence.
- These nodes with high degree compared to other nodes having lesser degree results in larger influence spread in the OSN.
- The Table below calculates the Degree Centrality for each of the 10 nodes given in the Figure below.

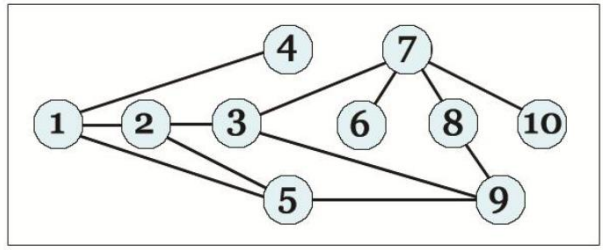


Fig. An online social network consisting of 10 nodes

Node v	1	2	3	4	5	6	7	8	9	10
Degree (v)	3	3	3	1	3	1	4	2	3	1
DC (v)	3	3	3	1	3	1	4	2	3	1

Table: Measurement of degree centrality (DC)

Common Influence Maximization Technique

- **The High-Degree Centrality Approach:**
- From the Table, we can check that node 7 has the highest DC score of four and can be considered to be the strongest networked node for spread of influence in the network, followed by nodes 1, 2, 3 5, and 9 as all are having a DC score of three.
- Hence, if we need to select a seed set of 5 nodes, the seed set would consist of nodes:

{7, 1, 2, 3, 5}

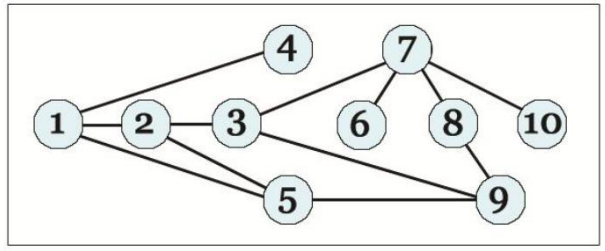


Fig. An online social network consisting of 10 nodes

Node v	1	2	3	4	5	6	7	8	9	10
Degree (v)	3	3	3	1	3	1	4	2	3	1
DC (v)	3	3	3	1	3	1	4	2	3	1

Table: Measurement of degree centrality (DC)

Standard Influence Maximization Techniques

- Two other standard methods of influence maximization include the following:
  - The *General Greedy* approach
  - The *DegGreedy* approach[2]

[2]***Reference***: Nandi, Gypsy, U Sharma, and A Das. "A novel hybrid approach for influence maximization in online social networks based on node neighborhoods." *Advances in Electronics, Communication and Computing*. Springer, Singapore, 2018. 509-520.

Text Analytics / Mining

Text Analytics

- **Text analytics** or **text mining** is the analysis of unstructured textual data using standard tools and techniques.
- For text analysis, the **unstructured input text** is at first **converted** into a **structured format** and then some typical text mining operations are carried out to generate the required output.
- Text analysis is very commonly used in social media analytics by considering as input the ***comments, tweets, reviews, discussions, emails, or feedbacks*** provided in social media by several online users.
- These social media contents are used for text analytics by using **linguistic, statistical, and machine learning techniques**.

Text Mining Tasks

- Some of the text mining tasks include text clustering, text categorization, document summarization, concept or keyword extraction, sentiment analysis, and entity/relation modeling.

- Let us now have a basic understanding of three of the common tasks carried out in text analytics, namely:
  - **Text categorization**
  - **Text summarization, and,**
  - **Sentiment analysis**

Text Categorization

- **Text categorization** or **text clustering** is the process of grouping textual data into clusters based on relevant categories.
- For example, if we are considering a set of documents that consists of news articles, it can be categorized into several clusters of documents based on the content, say sports, politics, fashion, and advertisements.
- This is shown in Figure as to how text categorization helps in classifying content into several groups.

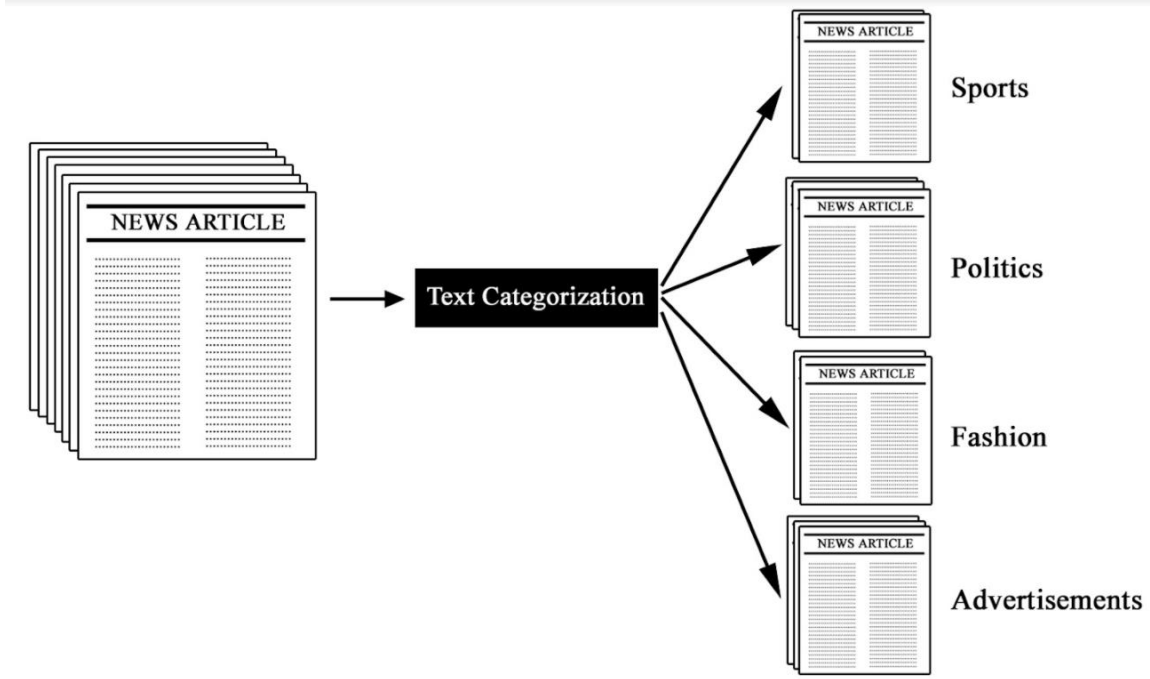


***Fig**: An example of text categorization*

Text Categorization

- For carrying out text categorization, any of these classification systems can be followed: – rule-based, machine learning-based, and hybrid.
- **Rule-based Text Classification**:
  - This approach of text classification categorizes text into organized clusters by using a **set of linguistic rules.**
  - Each such linguistic rule contains a consequent or predicted category (say, sports) and an antecedent or pattern.
  - The antecedent in rule-based text classification contains a series of conditions on the input and the consequent is like a label that defines the category.
  - The rules help in instructing the system to check the semantically relevant text elements and accordingly identify the relevant category.

Text Categorization

- **Machine learning-based Text Classification**:
  - While rule-based text classification relies on human crafted rules, machine learning-based text classification can be used by training machines to learn to make classifications based on past observations.
  - The training data contains labels which determines the category of output of a particular data. For the training, feature extraction is initially performed to transform the text into numerical values.
  - A vector is formed to represent the frequency of each word when compared with a predefined series of words.
  - Next, the machine learning algorithm is fed with training data that contains pairs of feature sets and labels or tags to produce a classification model.
  - It has been observed that machine learning-based text classification performs better than human crafted rule-based text classification.

Text Categorization

- **Hybrid Classification**:
  - As the name suggests, hybrid text classification combines both machine learning-based text classification as well as human crafted rule-based text classification to produce more accuracy in results.
  - Here, the base classifier is trained with the machine learning technique as well as a rule-based technique to be able to predict the output correctly, especially in case of conflicting tags.

Document or Text Summarization

- **Document or text summarization** is the technique of finding a precise summary of the bulk amount of data fed in as input to be able to interpret the key idea of the content of the text based on the generated summary.
- In short, this technique allows for creating a much-shortened version of lengthy documents.
- If done manually, the entire process of text summarization will be a very complex and tedious task, and sometimes maybe almost impossible within a short time frame.
- In such a case, one easy solution is to use machine learning algorithms that can be trained to identify the important sections of a document and accordingly produce a summary of the document.
- There are mainly two types of text summarization approaches followed in text analytics, namely the **extraction-based summarization approach** and the **abstraction-based summarization approach.**

Sentiment Analysis

- **Sentiment analysis** or **opinion mining** in social media analytics is the computational study of the opinions, emotions, and attitudes expressed by social media users toward a topic, an object, or an individual.
- Nowadays, social media gives immense importance to product reviews or topic reviews provided by online users which have led to the development and growth of blog repositories, discussion forum sites, and review sites.
- These sites allow online users to express their genuine views and opinions about a product or concept which can be positive, negative, or neutral review or comment.
- A prominent social media site often used for sentiment analysis is *Twitter* where users can tweet messages to express their opinion on any topic or person.

Trend Analytics

Trend Analytics

- **Trend analytics** mainly involves determining the possible drifts or trends over a period of time.
- Usually, historical trends are analyzed to determine future trends for a given phenomenon or feature. So, trend analytics is used **to predict future events**.
- The main task in trend analytics is comparing data stored during a period of time to analyze and visualize the change of trend in this data with time.
- For instance, **predicting the stock market trends** to allow users to understand and analyze in which company or organization it will be suitable to invest money.
- Trend analysis has proved to be very helpful in many applications, be it a study on **climatic change** or **stock market trends**.

Types of Trend Analysis

- There are three types of methods in trend analysis – temporal, geographic, and intuitive.
- Let us now have a basic understanding of each of the **three trend analysis methods**.
- **Temporal trend analysis:** Temporal trend analysis allows one to examine and model the change in the value of a feature or variable in a dataset over time.
- **Geographic trend analysis:** Geographic trend analysis is mainly involved in analyzing the trend of products, users or other elements within or across geographic locations.
- **Intuitive trend analysis:** This approach is more often used when there is a lack of large statistical data required to carry out trend analysis. Here, the data analyst need to behave like a futurist and based on logical explanations and study of behavioral patterns, a prediction of future trends is made. However, this method is prone to biases of the analyst and is comparatively difficult to analyze compared to the other trend analysis.