UNIT-III

EXTRACTION AND MINING COMMUNITIES IN WEB SOCIAL NETWORKS

* The extraction of web community utilizes web community chart A graph of communities.

* The main advantage of web community chart is existence of relevance between communities.

Notation use;

* $t_1, t_2 \ldots t_n$; Currently, a month is used as the unit time.

* $w(t_k)$; The web archive time at time t_k.

* $c(t_k)$; The web community chart time t_k.

* $c(t_k), d(t_k), e(t_k) \ldots$ Communities in $c(t_k)$

Types of changes:

* Emerge
* Dissolve
* Growth and Shrink

* Split
* Merge.

Evolution Metrics:

* Growth Rate
* Stability
  * Disappearance rate
  * Merge rate
  * Split Rate.

Other Metrics:

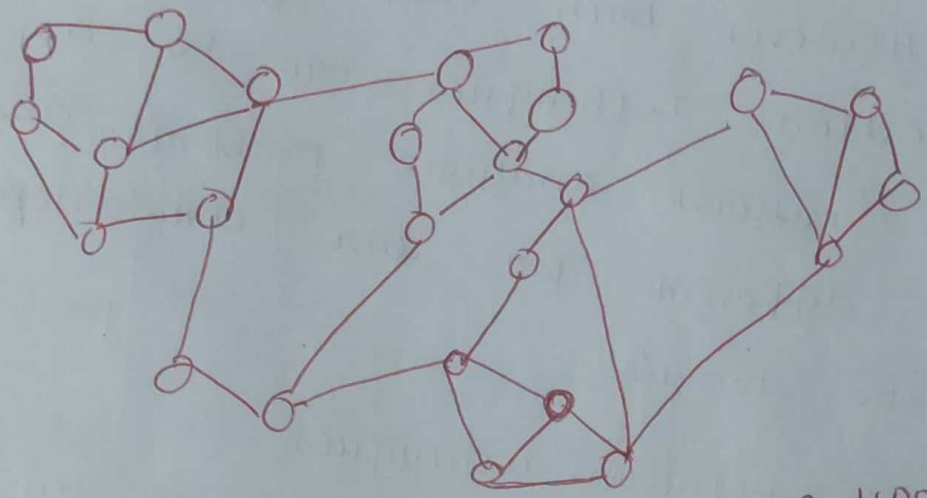
* Web Archives and Graphs.
* Split and Merged communities.
* Emerged and Dissolved communities.
* Growth Rate.

Detecting communities in social networks:-

* Communities will help us understand the structures of given social networks.
* communities are regarded as components of given social networks.
* communities will play Important roles, when we visualize large-scale

Social networks

* Relations of the communities clarify the processes of Information sharing and Information diffusions.



Many of your familiar with networks:

* Social media sites such as Facebook Instagram, Twitter etc.,

* Communities are a property of many networks in which a particular network may have multiple communities. Such that nodes inside a community are densely connected.

Why community Detection?

* Community detection can be used in Machine learning to detect groups with

similar properties.

Community Detection Vs clustering:

* clustering is a machine learning Techniques in which similar data points are grouped into the same cluster based on their attributes.

* However both clustering and community detection Techniques can be applied to many network analysis problems and may raise different pros and cons depending on the domain.

Community Detection Techniques:-

* community detection methods can be broadly categorized into two types.
  * Agglomerative Methods
  * Divisive Methods.

Agglomerative Methods:

* Edges are added one by one to a graph which only contains nodes.
* Edges are added from the stronger edge to the weaker edge.

* Divisive methods follow the opposite of agglomerative methods.

* In these, edges are removed one by one from a complete graph.

* There can be any number of communities in a given network and they can be varying sizes.

Types of communities:

1) URBAN

U - usually a large population so it can be noisy.

R - Residents can take buses or taxis to work.

B - Buildings are close by one another so sometimes people walk

A Apartments are a popular place to live so the buildings are all.

N - Night life is Busy because you can go to movies, theaters, or restaurants.

Definition of Community

* The word community Represents Subnetwork whose edges connecting inside of it are denser than the edge.

Types of communities: [continuation) (2) 2 mark

* Rural community
* Traditional community
* Solidarity community
* Urban community
* neighbourhood community.

Definition of Community:

* The word "Community" means a subnetwork whose edges connecting Inside of it are denser than the eages connecting outside of it.

Community can be classified into the following three categories;

1. Local definitions
2. Global definitions
3. Definition Based on vertex similarity

Local definitions:-

LAN; Local Area network

* The computers in a LAN connect to each other via TCP/IP ethernet or wi-fi.
* LAN is normally exclusive to an organization such as school, office, association (or) church.

Global definition;

Social networking platform; facebook, Linkedin, Myspace, orkut.

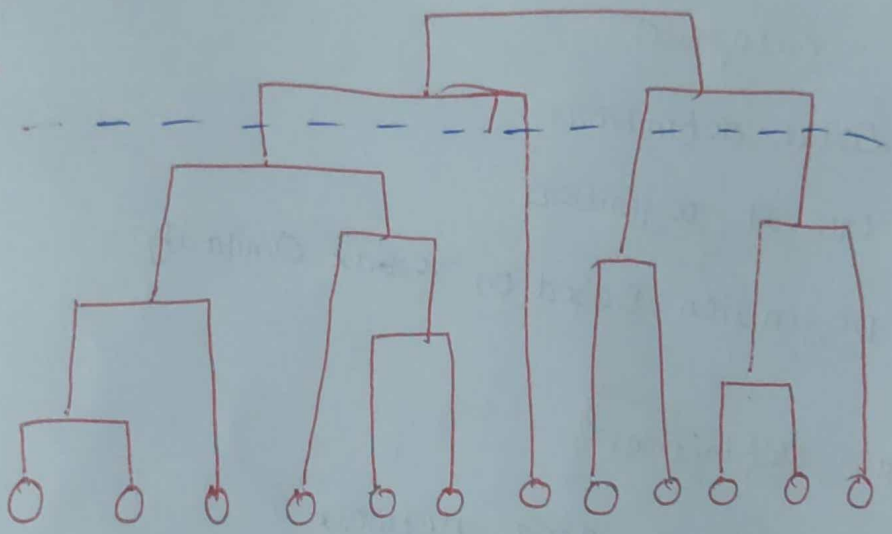
Microblogs; Twitter, GoogleBuzz.

Photo and video sharing; Flickr, youtube.

Definitions Based on vertex similarity.

* Communities are groups of vertices which are similar to each other
* some quantitative criterion is employed to evaluate the similarity between each pair of vertices.

Ex:



* This structure is called dendrogram, and highly similar vertices are connected in the lower part of the dendrogram
* Subtress obtained by cutting the dendrogram with horizontal line correspond to communities

Evaluating communities:-

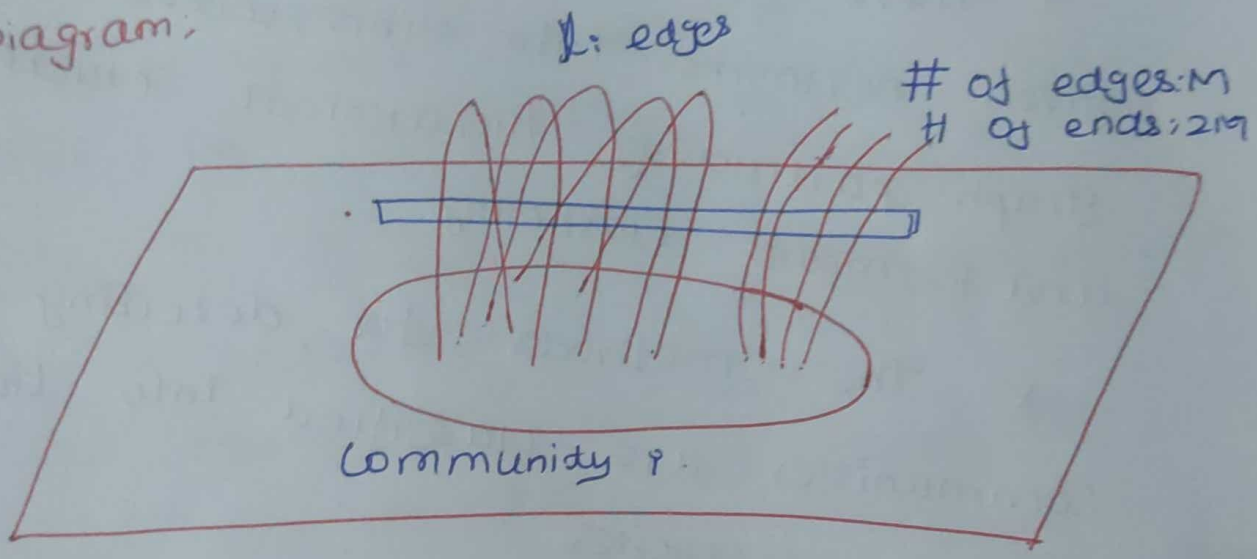
* Real community structure.
* quality function for evaluating how good partition is needed.

* The most popular quality function is modularity of newman and Girivan.

* The fraction of edges of the network inside the community

* Second Term Represents the expected fraction of edges.

Diagram:



$e_{ii} = \frac{Li}{M}$: Fraction of edges in community i.

$a_i^2 = \left[\frac{di}{2M}\right]^2$: fraction of edges when connected randomly.

* A subnetwork is a community if the number of edges inside it is larger than the expected in Modularity's null model.

* The Modularity of the whole network taken as a single community, is zero
* Modularity is always smaller than one, and it can be negative as well.

Methods for community detection and mining

* There are naive methods for dividing given networks into subnetworks, such as graph partitioning, hierarchical clustering and K-means clustering.
* The methods for detecting communities are classified into the following categories

1) Divisive algorithms
2) Modularity optimization.
3) Spectral algorithms
4) Other algorithms.

Divisive algorithms:

* A simple way to identify communities.
* The steps of the algorithm are as follows.

1) Computation of the centrality of all edges.
2) Removel of edge with largest centrality
3) Recalculation of centralities on the Running network
4) Iteration of the cycle from step (2)

* Edge betweeness is the number of shortest paths between all vertex pairs that run along the edge.

Modularity optimization.

* Modularity is a quality function for evaluvating partitions.
* This is the main idea for Modularity optimization.
* It has been proved that modularity optimization an NP hard problem.
* famous algorithms for modularity optimization is CNM [customer network management) algorithm.
* Another examples of the algorithm are greedy algorithm and simulated annealing

Spectral algorithms:

* Spectral algorithms are cut given network into pieces so that the number of edges to be cut will minimized.
* one of the Basic algorithm is spectral algorithm
* The Laplacian matrix L of a network
* All eigenvalues of L are real and non-negative.

other algorithms:-

* There are many other algorithms for detecting communities, such as the methods focusing on random walk, and the ones searching for overlapping cliques.

Applications of community mining algorithms:

* Network Reduction.
* Discovering scientific collaboration Groups from social networks.
* Mining communities from Distributed and Dynamic Networks.

Tools for detecting communities social network infrastructure and communities

* Several tools have been developed for detecting communities.

classified into the following categories.

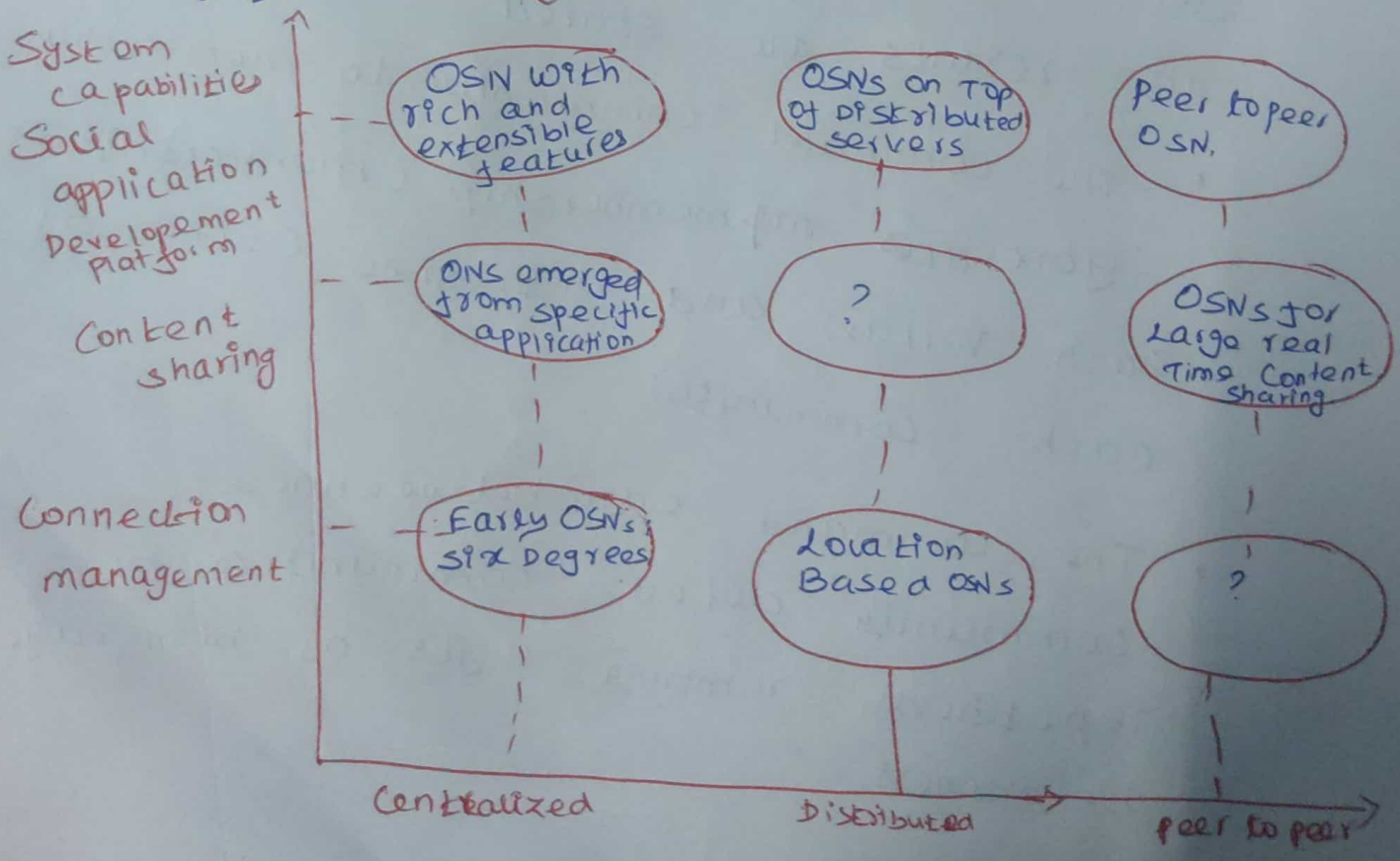
1. Detecting communities from large scale network.
2. Interactively analyzing communities from small networks.

* Graph modelling language (GML) is one of the formats for representing networks

- The common fastgreedy- community is for maximizing modularity greedily, and its results are stored in variable gr.
- The command community.to.membership: generates membership (membership for each vertex) and membership size (size of each community).
- The command edge.betweenness: community detect communities by repetitively removing edges of high edge betweeness.

Decentralized online social networks

* Online social network (OSN) is an online platform that
  - provides services for a user to build a public profile and to explicitly declare the connection between his or her profile with those of the other users.
  - enables a user to share information and content with the chosen users or public.
  - supports the developement and usage of social applications.
  - user can Interact and collaborate with both friends and strangers.

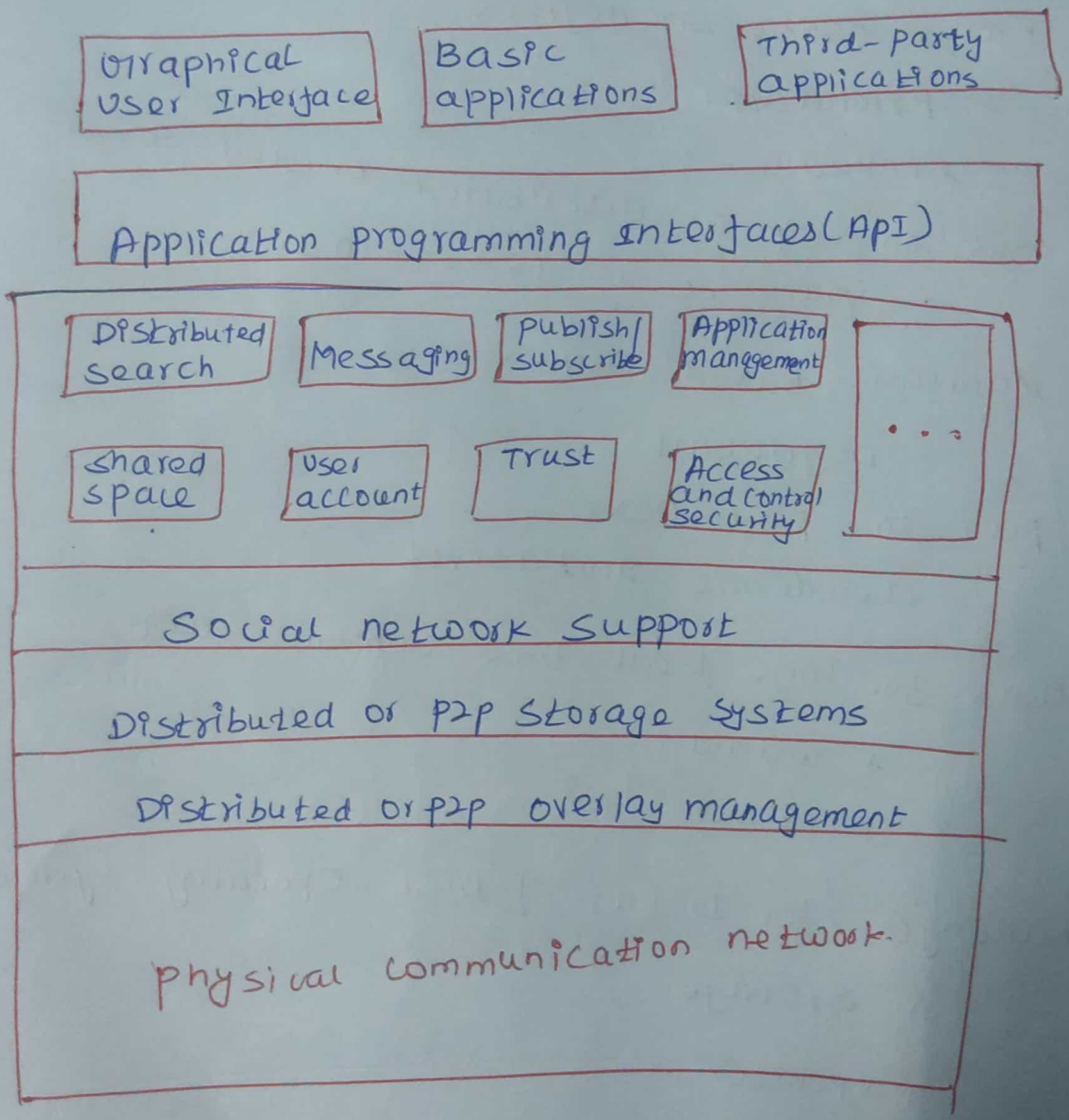


* OSNs with Rich and extensive features
  - Face Book
  - My Space.
* OSN emerged from specific applications:
  - Flickr
  - YouTube
  - Twitter.
* Early OSNs: six degrees,
  - Friendster
  - Linkedin
* OSNs on top of distributed trusted servers:
  - young et al. (2009, W3C).
* Location Based OSNs:-
  - Dogdeball.
* peer_to peer OSNs:-
  - academic initiatives.
* OSNs for large Real-time Content sharing.
  * Wuala.
  * Tribler.

challanges for DOSN: [DISK Operating system].
  * Storage
  * updates
  * topology
  * search, Addressing.

* Openness to New Applications
* security
* Robustness
* Limited peers
* Locality

General purpose DOSNs. [Disk operating system].



* The Reference Architecture consists of six layers.
* Lower layer of this architecture is the Physical communication network.
* The Distributed or P2p overlay management provides core functionalities to manage Resources in the supporting Infrastructure of the system.
* on the top of this overlay is the Decentralized data management layer.
  * which Implements functionalities of a distributed or peer-to peer Information system to query, insert, and upaate various presistent object to the systems.
  * the social networking layer Implements all basic functionalities.
  * Social networking layer exposes and Implements an application programming interface. To support pevelopement of new applications
  * The top layer of the architecture Includes the user Interface to the system and various applications.

* The of to Top of the development platform provided by the DOSN

Multi-Relational characterization of dynamic social network communities:

* The characterization of communities in online social media is presented using computational approaches grounded on the observation from social science.

* Motivation: human community as meaning-making eco-system:- The semantic is an emergent artifact of human activity.

* Human activity is mostly social, and social networks.

motivating applications:

* Real human network communities.

Example Applications Include:

* Context-Sensitive information search and recommendation.

* Content organization, Tracking and monitoring.

Data characteristics and challenges:

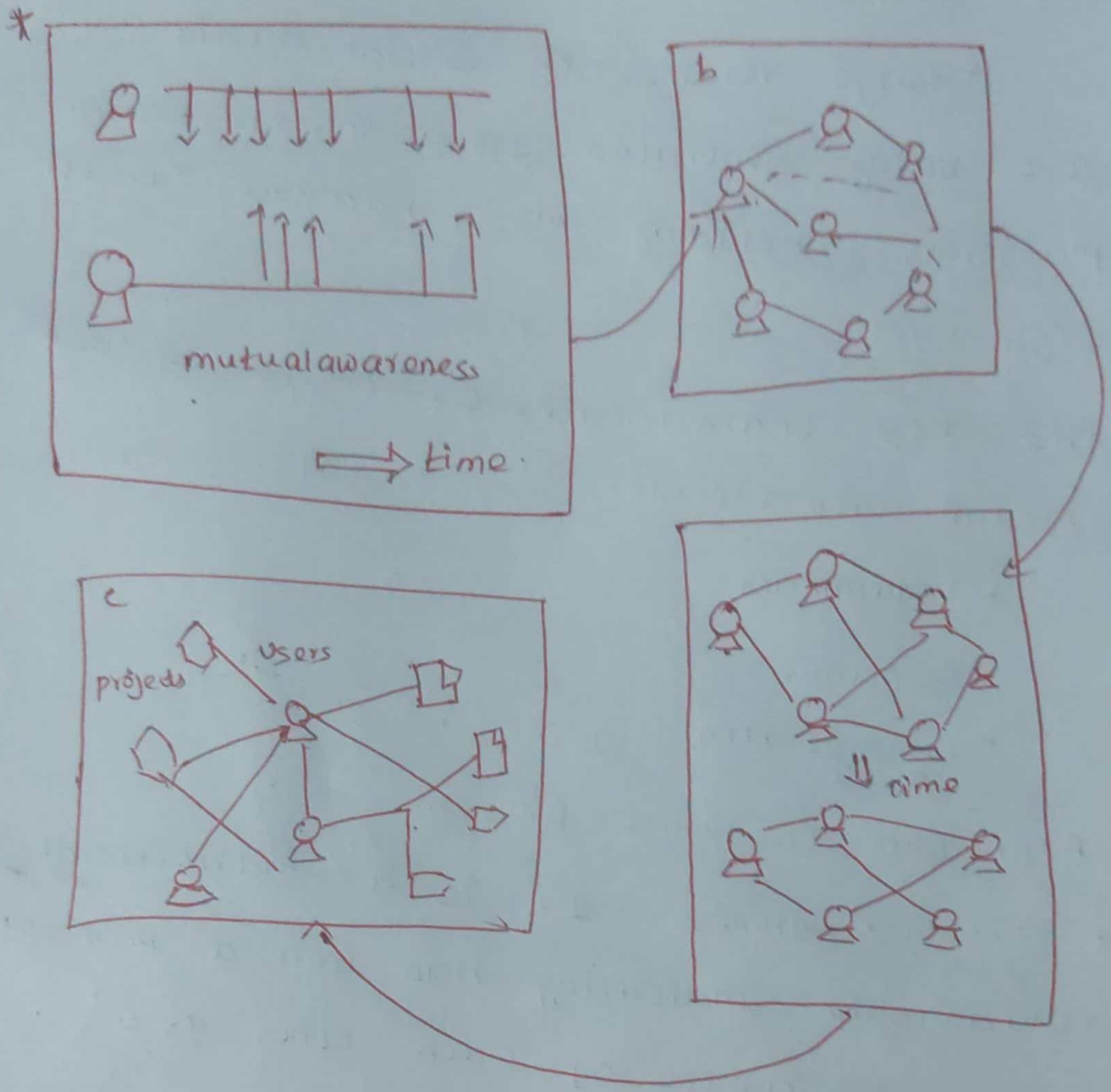
*Large volumes of social media data are being generated from various social media platforms including blogs, FaceBook, Twitter, Digg, Flickr

The key characteristics of online social Media data include:-

* Voluminous.
* Dynamic
* Context-rich.

Approaches to three problems:-

* Mutal awareness:- It is a bidirectional relationship indicating how well a pair of bloggers is aware of each other, as a
* Facet Net: The community structure at a given timestep is determined both by the observed networked data.
* Metafac: Metafac is the first graph-Based tensor factorization framework for analyzing the dynamics of heterogeneous social networks.


mutual awareness
time
b
c
projects
users
time