Trace Analysis, Clustering and Network Theory

NOMADS

Usage Trace in RAW Form

- 2007 11 17 00:00:03 EST b430bar-win-ap1200-1 [info] 189341: Nov 17 00:00:02 EST: %DOT11-6-ASSOC: Interface Dot11Radio0, Station 001b.fcb2.4fc9 Reassociated KEY_MGMT[NONE]

- 2007 11 17 00:00:19 EST elm-authgw-bs2100-1 [notice] user_tracking: event=user_login_successful&loglevel=notice&obj=user&ipaddr=10.249.52.174&name=bob&msg=Login RADIUS user dung on Primary RADIUS server at [00:1b:fc:b2:4f:c9]/10.249.52.174 as role Authenticated, login time = 2007-11-17 00:00:19, sessionID = 00:0E:0C:33:08:BA:1195275619616 45&

Information Extraction

- HOST_MAC VARCHAR2(20)
- START_TIME DATE
- END_TIME DATE
- AAP_NAME VARCHAR2(150)
- START_TIMESTAMP NUMBER(38)
- END_TIMESTAMP NUMBER(38)
- ATIMEZONE VARCHAR2(20)
- RECORD_TYPE NUMBER(38)
- ATRANSACTION_ID NUMBER(38)
- AAP_BLDG VARCHAR2(20)
- DAP_BLDG VARCHAR2(20)
- DAP_NAME VARCHAR2(150)
- DTIMEZONE VARCHAR2(20)
- DTRANSACTION_ID NUMBER(38)
- ROAM_MAC VARCHAR2(20)
- Udayan Mapping with UF Phone Directory
- Number Theory and Counting

My Experience → Logic Comes First

- Focus on the Idea and NOT on the Data
- Selection of right data
- Selection of right Format
- Data Representation should be correct
- Start with a small Sample (training set) and move

Most Common Form

Start Time	Location	Duration
8175587	172.16.8.242_11006	5284
8182291	172.16.8.242_11006	14463
8243584	172.16.8.242_11006	12369
8256573	172.16.8.245_31031	20853
8283387	172.16.8.242_11006	13545

Clustering in Spatio-Temporal Environment

Clustering

- Problem:
  - To get
    - List of users online together
    - Cluster size
    - How many times?
    - List of all locations, time etc.

- Challenges
  - Each User has ~100+ sessions
  - There are ~12000 Users
  - *Only ~5 years to finish your PhD*

Basics

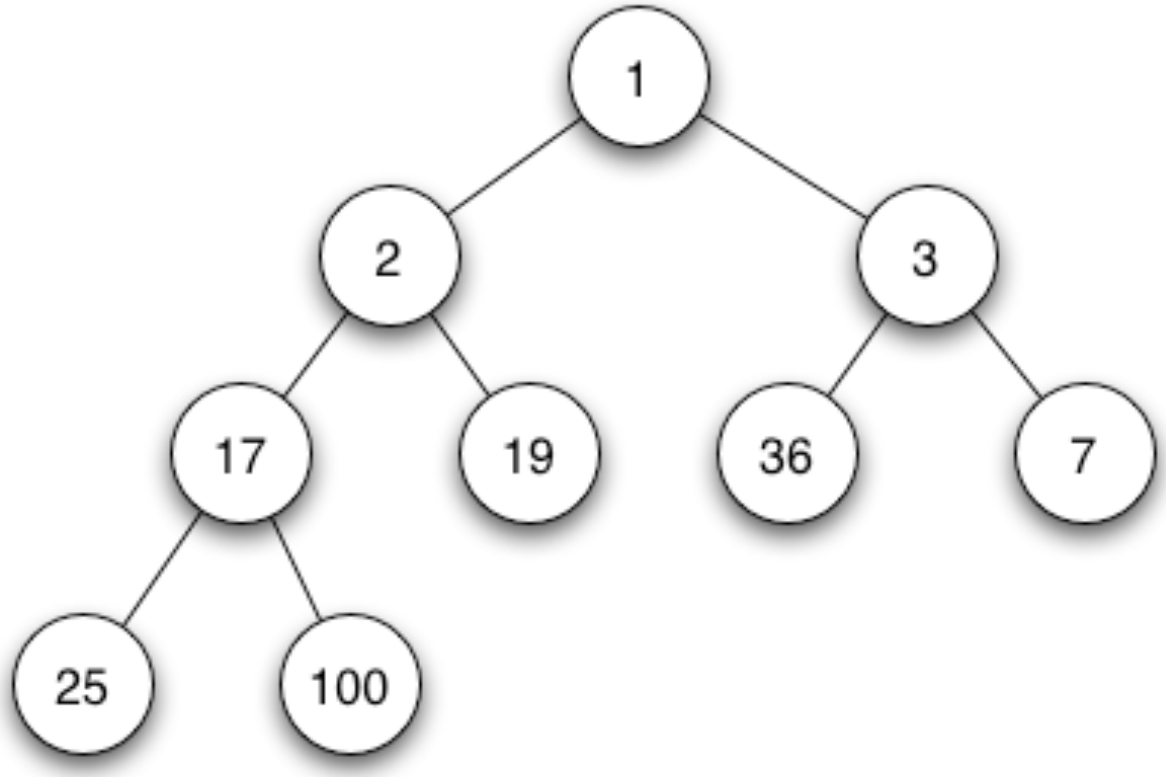
- Rule#1: Users connected to same Access Point
  - if(user_x.location.equals(user_y.location)){
  - }
- Rule#2: Users have intersecting time intervals
  - if(user_x.st <= user_y.et && user_x.et >= user_y.st){
  - }

Intersecting Intervals

User A: -------- -------- -------- --------

User B: -------- -------- -------- ----

Min Heap



Sample Output

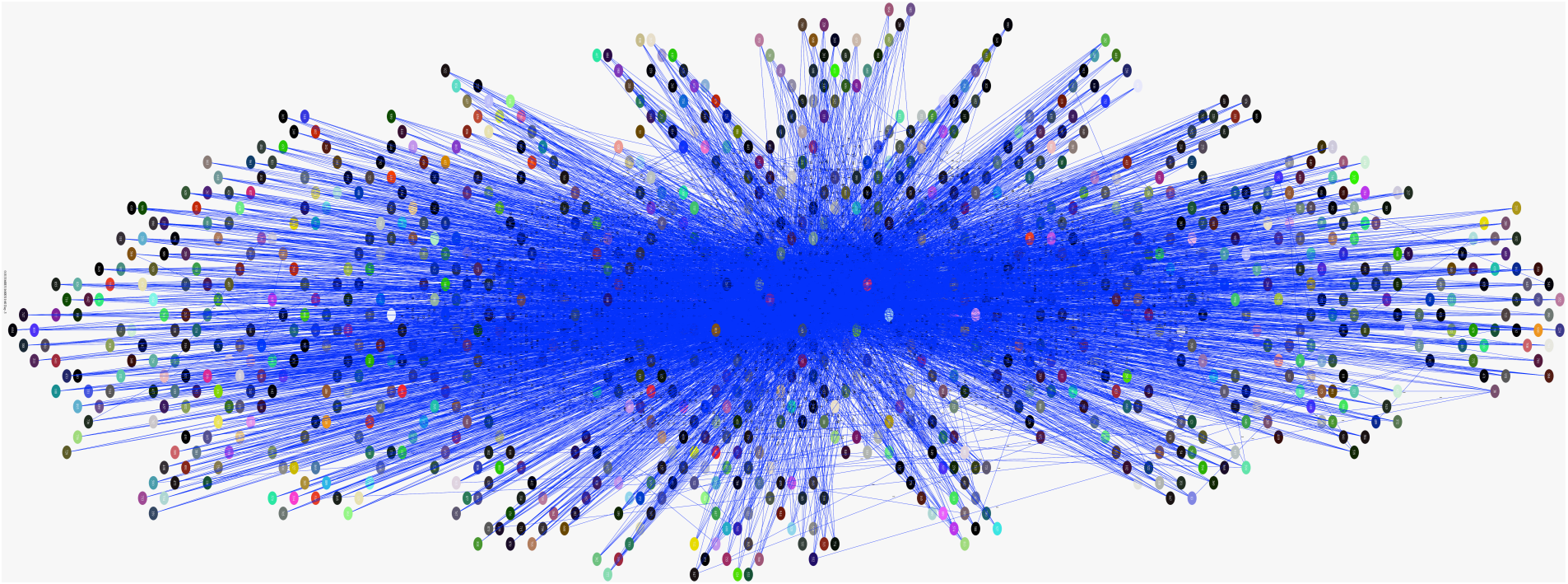
- 2 6 14114 14988 17317 24220 29687 29710
- 2 3 5727 16499 24623
- 1 3 17798 21831 26081
- 12 5 15659 15869 19035 24539 24729
- 4 3 9650 13673 15720

Possible Extensions

- Maintain a HeapMap / Vector of locations
- Maintain a timestamp schedule they met
- Maintain average time they met
- Regularity in meeting with this group or other group
  - FFT (Yibin, Sungwook)
- Periodicity of meeting etc.
  - Markov Chain Markov Models (Jeeyoung)

Min Heap Implementation

- First Sort using Location
- Then Sort using start time
- User minHeap to extract the earliest finish using "End Time"
- Insert into minHeap using "Start Time"
- Add all extracted nodes into the HeapMap as key.
- For every extraction, check if the same set present in HeapMap, if yes, increment the value by one.



Measures and Metrics

An introduction to some standard measures and metrics for quantifying network structure.

Measures and Metrics to look

- Degree Centrality
- Eigen Vector Centrality
- ~~Katz Centrality~~
- ~~Page Rank~~
- Hubs and Authorities
- Closeness Centrality
- Betweenness Centrality
- Groups of nodes (cliques, plexes, cores and components)
- Transitivity
- ~~Reciprocity~~
- ~~Signed Edges and Structural balance~~
- Similarity
- Homophily and Assortative mixing

Background

- If we know the network structure, we can model and calculate from it a variety of useful quantities or measures that capture *Particular Features* of the network topology.

- These measures in turn helps us to design new protocols, data transmission, adaptivity in the network and other overheads (one step further)

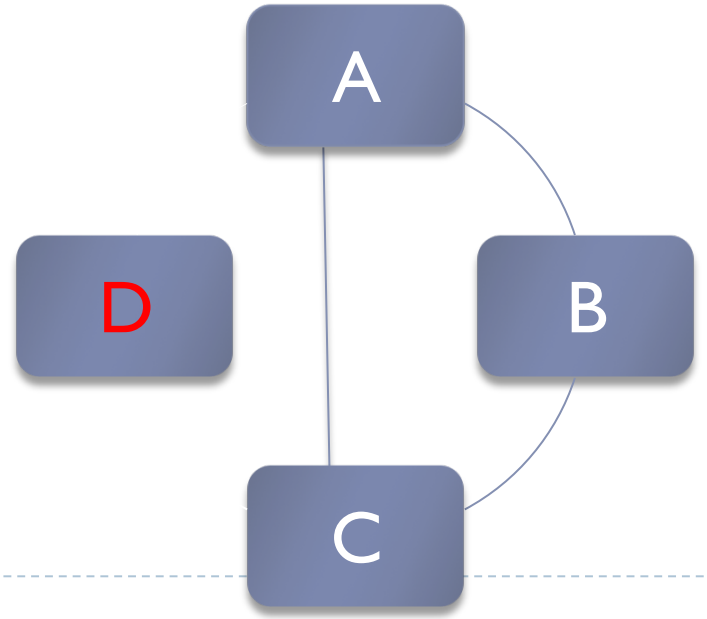
Background…

- To generate a Graph G(N,E) with *N* as a set of vertices and *E* as a set of edges that connect *N* nodes.

- When two or more nodes in WLAN trace have intersecting *Online Time* and *Location*, I assume there exists an Edge between those nodes.
  - Location: Access Point, Building, group of Access Point etc.

- Any other idea?

Node ID	Start Time	End Time
A	10001	12000
B	9000	11000
C	10000	11989
D	5000	80000



Centrality

- Who are the most central or important nodes in a network?
  - A set of Access Points
  - A particular building
  - A set of users
  - Websites
  - Period of a day or a week, pathways, etc.

- Helps to know:
  - Popular spots, overhead, efficiency and deployments of proper capacity channels, routing efficiency

Degree Centrality

- Degree of a Node: the number of edges connected to it.
  - Undirected case: only one type of degree
  - Directed case: in-degree, out-degree.

- Helps to know:
  - Influence, more access to information, critical for network connectivity.
  - Location analysis: Location with high user activity can have better evacuation routes, designing back up system in case of any emergency.

Eigen Vector Centrality [EVC]

- A node's importance is increased by having connections to other nodes who themselves are important.
- Eigen Vector centrality gives each node a score proportional to the sum of those of its neighbors.
- Any idea, where to use it?

Eigen Vector Centrality [EVC]…

- EVC can be *large* because a node has *many neighbors* or it has *important neighbors* (or both)

- Helps to know:
  - A set of critical locations,
  - relationships among AP,
  - Can be used in trust,
  - page rank in WWW,

Hubs and Authorities

- A node has high centrality if those that point to it have high centrality.
- Example:
  - Locations to find specific information
- In some cases, its appropriate to accord a node high centrality if it points to other with high centrality.

Hubs and Authorities…

- Two important type of Nodes:-
  - Authorities: nodes with useful information
  - Hubs: nodes that tell us where the best authorities are to be found

- Example: In Emergency situation where Evacuation Routes can be made to reach shelter.

- More info:
  - HITS Algorithm: Hyperlink induced topic search, by Kleinberg

Closeness Centrality

- This measures mean distance from a node to other node.

- Let d_{ij} is the length of a geodesic path from *i* to *j*, meaning the *number of edges* along the path. Then mean geodesic distance is averaged over all nodes (*n*) is

$$l_i = \frac{1}{n} \sum_j d_{ij}$$

Closeness Centrality…

- Take low value, if separated from others by only a short geodesic distance on average.

- Better access to information, or more direct influence,
  - In Small world effect
  - Information localization
  - Virus Spreading

Betweenness Centrality

- Measures the extent to which a node *lies on paths* between other nodes.
- Instead, it measures how much a node falls "*Between*" others.
- Example: internet routing, DTN message transfer

Betweenness Centrality…

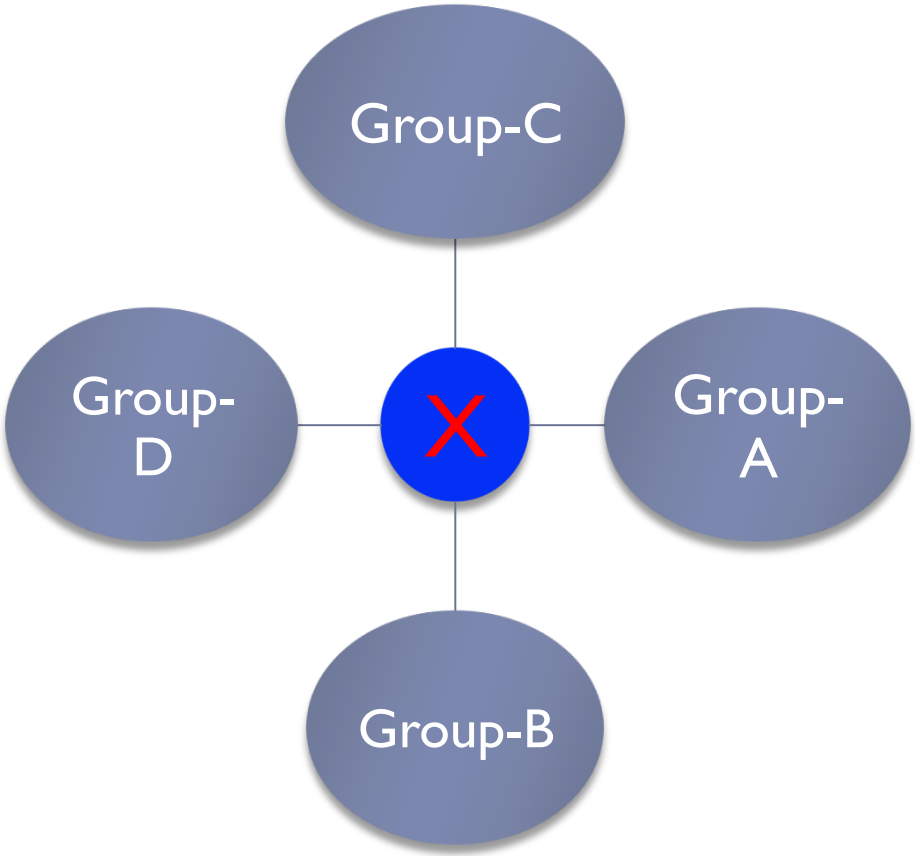
- Nodes with higher Betweenness have considerable influence within network for *a control over information* passing between others.

- Removal disrupts the major communication between other nodes, they lie on major paths.

Betweenness Centrality…

- Low degree node with high Betweenness

- Example:
  - Maximum flow of information
  - Potential location for disruption of the network



Comments and Discussion

Groups of Nodes

- Many networks divide naturally into groups.
- The definition and analysis of groups within networks is a large and fruitful area of network theory.
- cliques, plexes, cores and components

Cliques

- Maximal subsets of the nodes connected to each other.
- Occurrence in an otherwise sparse network is normally indicate highly cohesive cluster.
- Examples:
  - Nodes belong to same location, high cohesive ones.

Plexes

- Cliques: So much Stringent
- *k-plex*: a *k-plex* of size *n* is a maximal subset of *n* nodes within a network such that each node is connected to at least *n-k* of the others.
- Useful: in discovering groups in the network,
- *Example: create a community such that each node has at least n-k neighbors.*
- No defined values, start from small values

Cores

- A *k-core* is a maximal subset of nodes such that each is connected to at least *k* others in the subset.
- For practical reason, easier to compute than plexes,
- *K-core* of *n* nodes is also (*n-k*) plex.

Components and k-Components

- A maximal subset of nodes such that each is reachable by some path from each of the others.

- Useful: k-component, a maximal subset of nodes such that each is reachable from each of the others by at *least k node-independent paths.*

- In data network, number of independent path between two nodes or locations is also independent, routes with *data, information transfer* might take.

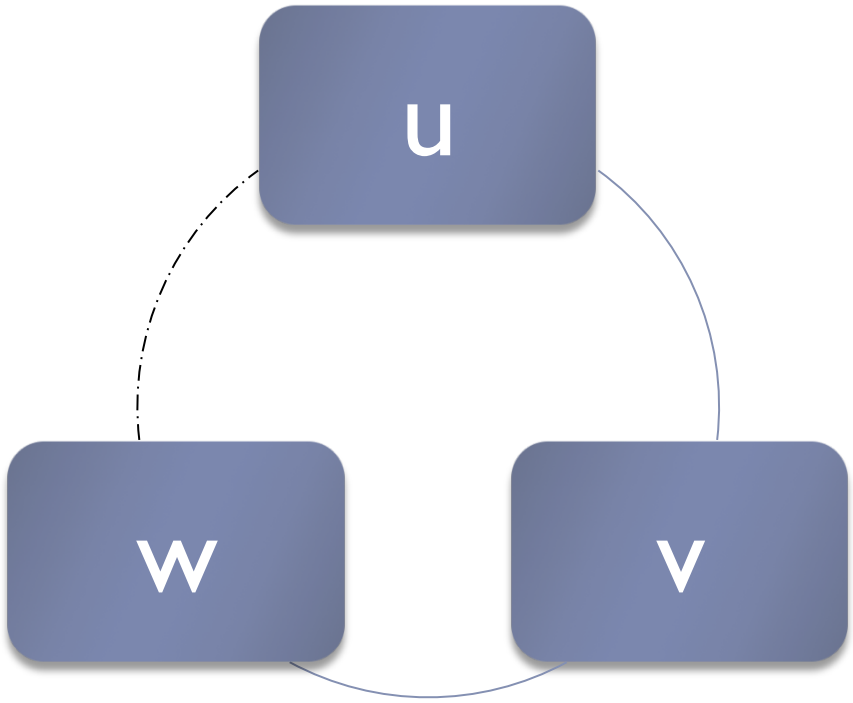
Transitivity

- "the friend of my friend is also my friend"
- In a network graph, "edge" creates a transitive relation
- Good for:
  - Knowing explicit vs. implicit network relationship
  - u → v,
  - v →
    - w, ➔ u → w ??
  - May or may not

Clustering Coefficient [CC]

- Fraction of paths of length two in the network that are closed.

- The path uvw (solid) is said to be closed if the third edge directly from u to w is present (dashed)

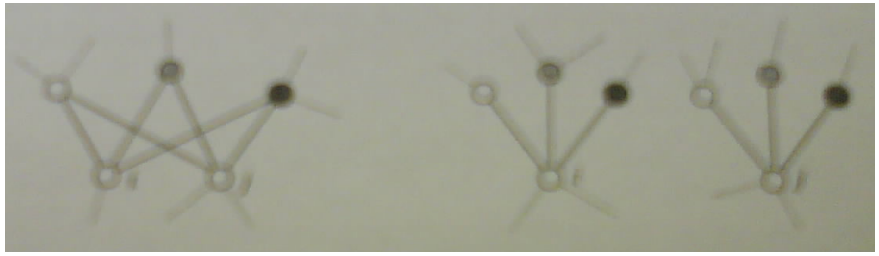


Clustering Coefficient [CC] …

- That is, count all paths of length two and count how many of them are closed, and divide the second number by the first to get CC in the range 0 - 1.
  - CC = (number of closed paths of length two) / (number of paths of length two)
- But can always be extended for more than two.

Similarity

- Which node u is most similar to node v?
  - And in what sense?
- Two types of similarity
  - Structural Equivalence
    - Sharing same neighbors
  - Regular equivalence
    - Two history student in two different universities
- Profile-Cast



Cosine Similarity

- To check structural equivalence would be just a *count of number of common neighbors* two nodes have.

- Inner dot product of two vectors, the number of common neighbors of the two nodes divided by the geometric mean of their degrees.

- Values ranging from 0 – 1

Homophily and Assortative Mixing [By Degree]

- High degree node connected to other high degree node, and the low to low.
- Can have Disassortative mixing also, indeed interesting fact to explore. Any one tried?
- High with high:
  - To get a core of such high degree nodes surrounded by less dense nodes,
  - Example: backbone routers.

Homophily and Assortative Mixing [By Degree]

- Disassortatively mixed by degree, high degree nodes tend to connect low degree nodes, *creating a star like* features in the network.

Comments and Discussion