Frequently asked questions & Presentation schedule

How to schedule my presentation in the seminar

Fill in your choice at the following shared google document https://1drv.ms/w/s!AtcJs3OTsMZuiR5B-CMDHQqon-NH

How to visualize large graph with thousands of nodes?

The idea is not to visualize all the nodes of the network on the same page and copy and past this on your report, that is not interesting and not doable as we want to highlight the main components and interesting nodes of the network. You can screw up, down, left and right to identify this when using basic NetworkX plot. Alternatively, you may use more advanced graph library that you can import to your networkX. For instance, Graphviz, PyGraphvis, pydot, can provide some more interesting visualization that highlights such interesting components and nodes..

How to use k-plex in NetworkX?

k-plex is not directly implemented in NetowrkX

You can look at alternative existing implementations elsewhere.. For instance in https://github.com/bachsh/kplex/blob/master/tests.py (need to check if it is updated)- or another source

You can also make some program on your own.. For instance you can start by identifying the largest component in the graph, say V, and the smallest degree d of all nodes of V, then the k-value (in k-plex) whose outcome will be the subgraph V is |V|-d. You can proceed by exploring other components in the graph and identify the corresponding value of k accordingly as previously.. Check for specific value of k, the result should be associated to the largest set (subgraph)..

How to use power of law fitting and its confidence?

Once you generated the points in log scales, you can use any library with linear fitting with confidence value.. For instance, you can use SciKit library, with linreg.predict function, you can select a confidence level you want, usually, 95% and trace the two lines corresponding to the 95% confidence and see if all points fall within this interval.. You can also find similar functions in statsmodels package..

The tweets and follower Id take long to generate?

We are aware of the limitation of the Twitter API, and you may not be able to achieve the number of connections set in the project specification. Feel free to make some assumptions to restrict the size and scope but must ensure to have some connections.. otherwise you will end up with fully disconnected graph which is not interesting.. In the worst case scenario where you cannot generate followers based network and you end up with a fully disconnected graph, you can switch to another interpretation of edge.. For instance, use the amount of textual overlapping of the two tweets when exceeding a certain threshold to decide on the establishment of edge between the two nodes as defined in some other projects (try to imitate such reasoning as defined in that specification)..

The search of communities takes long?

It is true that running Girvan-Neuman takes time, feel free to explore other already implemented community detection algorithms in NetworkX and provide some rational quantifiable justification for your choice. Besides, it requires prior knowledge about k (number of partitions). For this you can save time by trying to visualize your network first to make an idea about the different partitions, so select k accordingly, then you may run Girvan-Neuman for k, k-1, k+1, and calculate quality index (partition performance and/or coverage) for k, k-1, k+1 to show that it provides better quality index for k. You can also try less computationally expensive algorithm available in NetworkX (e.g., label propagation, fluid community detection) or elsewhere, i.e., Louvain's algorithm (https://perso.crans.org/aynaud/communities/) or any other you identified yourself.. Just justify through a small comparison to show its merits with respect to the one mentioned in the specification.

How to calculate the diameter and clustering for the whole graph with NetworkX?

It is true that NetworkX function for calculus of diameter assume a connected graph so you have to use first NetworkX to identify the various components, and then calculate the diameter of each of these and then take the maximum value.. Similarly, the same for clustering coefficient.. it is easy to use calculate the local clustering coefficient of each node and then take the average.. You may also use the function transitivity in NetworkX, which provides global clustering coefficients according to number of triangles and triplets identified.

It is an error to endup with trivial values of infinity for diameter or clustering coefficient in case of error message by NetworkX functions..

How to draw cumulative distribution?

The cumulative distribution is different from standard distribution we have seen in handout for degree distribution for instance.. Cumulative distribution is always a decreasing function..

For cumulative degree distribution.. Assume for instance, you n1 nodes with degree 1, n2 nodes with degree equal to 2, n3 nodes whose degree is equal to 1, …, nk nodes whose degree is equal to k. Then the your start by calculating the total number of nodes is greater or equal to 1 (which is A1=n1+n2+..+nk). Similarly, number of nodes whose degree or equal to 2 is equal to A2=(n2+n3+..nk), Number of nodes whose degree is k is Ak=nk. Then you draw the graph by fitting the points (1,A1), (2,A2), (3,A3),..(k,Ak).. and use a fitting function of your choice – can be just linear fit.. or other.. and you will see a decreasing function..

Same thing applies for cumulative clustering coefficient.. where you calculate the local clustering coefficient xi for each node, rearrange these values in decreasing order.. and repeat the above process when you sum up for each xi to count the number of nodes whose local clustering coefficient is greater or equal to xi. This process is similar to the concept of cumulative distribution in probability so you can see tutorial in statsmodel package.. if you do not want to program this yourself..

How to handle all specifications given the timetable?

As already pointed out in the guideline document, the pass mark for each project is that 50% of project specification is achieved. So, you should highlight this achievement.

How the seminar presentation will be counted toward the total mark?

The presentation is worth 30% of total and 70% by report