# MAS2, Analysis of a larger (real) network

## US patents

### Number of nodes and edges



### Average In/Out degree



### Average Clustering Coefficient



### Most cited patents



## Yahoo messages

This network represents users of application Yahoo and their interaction with each other – how they send messages to other users. This network is dynamic, because each sent message has also its timestamp. I thought of the edges in this network as undirected. Meaning of the edge = communication was established between two users. It is not important who started the communication with sending message. Edges can repeat – In that case I increment the weight of the edge. It means that the nodes were communicating with each other more then one time.

**Constructed network is undirected, weighted and temporal.**

In total there is **100 001 actors**.

I preprocessed this network to 25 timestamps. Each timestamp contains all previous data + new data from current timestamp. This way i was able to construct temporal network. In this network we don’t talk about timesteps like years,day,minutes,… but just as an abstract timestamp. We would need to contact authors to get more information about real facts about this data.

### Number of nodes and edges



This graph shows number of nodes and edges in log scale. We see that nodes reach limit and only new edges start to appear.

### Average degree



On this graph we can see that the weight of existing edges is increasing because we are adding edges, where they already are. So instead of that we increase the weight. The weight in this network means how much nodes communicated with each other. We include repetitive messages. We can also see that after some point we are adding only edges and the number of nodes stagnates.

### Average Clustering Coefficient



In the dynamic network we can see that the Clustering Coefficient doesn’t ever reach values above 0.22. We can assume that in this network exists a lot of connected components. And also a lot of closed communities, that aren’t connected to the rest of the network. We definitely cant say that this network contains only one connected component.

### Number of connected components

At the last timestep of this dynamic network there is only **360** connected components. But at the beginning of this network – at the timestep 1, there is **2718** connected components. **We definitely can say that during the timesteps the number of nodes stopped increasing, but the number of edges was still increasing and thanks to that the number of connected components was reduced.**

### Nodes with max degree



We can see that already from the timestep 4 the node with ID 27278 started to dominate with the number of links which it has. We can assume that this node is probably some important person who is chatting with all other.

### Density of network



### Edge prediction

In this experiment I tried to use first timestep for prediction of edges in the second timestep (layer 0 -> layer 1). Predictions were done only on small sample of nodes (5 000 – every time the same nodes), because of complexity of the network.

* **CN** – Common Neighbors
* JC – Jaccard Coefficient
* **AA** - Adamic/Adar Index
* **PA** - Preferential Attachment
* **RAI** - Resource Allocation Index
* CS – Cosine Similarity or Salton Index
* SI – Sorensen Index

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Predicted Edges | Sensitivity | Recall | Specificity | Precision | Fallout | Accuracy |
| PA,10 | 5937287 | 0.9766 | 0.9766 | 0.3981 | 0.0022 | 0.6019 | 0.3989 |
| PA,1000 | 934 | 0.8913 | 0.8913 | 0.9997 | 0.8437 | 0.0003 | 0.9995 |
| AA,0.25 | 12749 | 0.9514 | 0.9514 | 0.9954 | 0.3024 | 0.0046 | 0.9953 |
| AA,0.5 | 7040 | 0.9361 | 0.9361 | 0.9975 | 0.4376 | 0.0025 | 0.9973 |
| AA,0.75 | 4301 | 0.9295 | 0.9295 | 0.9985 | 0.5627 | 0.0015 | 0.9984 |
| RAI,0.25 | 3400 | 0.9178 | 0.9178 | 0.9988 | 0.6154 | 0.0012 | 0.9987 |
| RAI,0.5 | 1859 | 0.9009 | 0.9009 | 0.9993 | 0.7378 | 0.0007 | 0.9991 |
| RAI,0.75 | 1460 | 0.8959 | 0.8959 | 0.9995 | 0.7795 | 0.0005 | 0.9993 |
| CN,2 | 2420 | 0.9164 | 0.9164 | 0.9991 | 0.6947 | 0.0009 | 0.9989 |
| CN,10 | 187 | 0.8916 | 0.8916 | 0.9999 | 0.9647 | 0.0001 | 0.9997 |

It can be seen that preferential attachment model did not give us good results as we chose too low a threshold and hence did not predict any edges in this experiment. For Preferential Attachment model we need to choose much more higher thresholds. When we choose threshold = 10 for Common Neighbour model we see that we are adding new edges only if these nodes have at least 10 common neighbours. For this model we should set the threshold according to average degree common neighbours metric.

The Adamic/Adar index and Resource Allocation Index gave us much better results. As already seen in the exercise tasks. By setting the threshold value in the interval <0.0f, 1.0f> we can easily influence the creation of new edges.

**When we take into account the meaning of this network, then in this experiment we tried to predict who would start to text with who in the future.**

### Simulation of influence spread

In this part I tried to simulate influence spreading through a network using algorithm - Independent cascade (IC) propagation model. I made this experiment on first and last timestep in network, to see huge difference in the connectivity of the network at the last timestep vs at the first step. From this we can also assume, that during the time the network gained more and more connections.

I made this experiment for 2/4/10 number of initial nodes. So three times I choose using my heuristic best 2/4/10 candidates. I was choosing candidates based on their value of clustering coefficient and degree. I tried to find such node with biggest degree and lowest clustering coefficient. These nodes act like a hubs in a real network. **In this heuristic approach I performed 2-level filtering of candidates.** First I chose candidates from whole network and then I performed another filtering on this candidates.

#### First timestep vs Last timestep



We can see that at the last timestep of dynamic network the network is much more dense. It is much more easier to spread influence because of the amount of existing edges, that yet weren’t present in the first timestep.

I chose parameter p=0.25 as value of probability of infecting node.

In candidates there is often a node with id **41159, 23189, 11677**.

According to the nature of the network, we can infer that these are nodes (people) who have been communicating frequently in this network (sending Yahoo messages).

### Conclusion

This network was fine example of temporal social network. We could see how new nodes appear and how new edges connects communities. I would definitely say that this network describes how communication between users of application Yahoo was ongoing during some time period.