Team members: Dimitar Kochev, Matei Pop, Mihael Stoyanov, Lorand Ladnai

Our data: [Email-EU](https://networkrepository.com/ia-email-EU.php?fbclid=IwY2xjawGUDyVleHRuA2FlbQIxMAABHfhin88U-90y_9Nj6tjd_D9ti7egzabl5PlBdbu_nV1lQnGrYGhFbt9n9Q_aem_z1qhvK7ifxaDvxHungKAOA)

Description: Our dataset contains a synthetic network generated from email data of a European research institution. It is considered a sparse network due to the fact that the number of edges is relatively smaller than the amount of all possible connections. Each node is an individual entity and the edge signifies a communication between two people. The vertices have no attributes. Because we assume a communication to be mutual the graph is undirected. Edges are also unweighted.

Question: How do you detect misinformation in a communication network? Can one prevent the spread?

Research scenarios:

* Case 1: Misinformation starts from a highly connected node.
* Case 2: Misinformation starts from a node with a small degree.
* Case 3: Misinformation starts from a random node.

Possible tools for the analysis: SI model; SIR model (who should do fact checks?); Random Walk;

Network stats:

Questions for session 1:

1. Do the cases we want to research represent “network-aware” models?
2. Should we drop one of our cases?
3. What model would be best? SIS, SIR or SEIR? Do we want to test different models with the different cases.
4. How we can define a highly connected node in the context of our graph?
5. How long should our presentation be?

Fact-Checking Behaviors

* In 2021, 47% of people aged 16-74 years in the EU encountered untrue or doubtful information on news websites or social media during a 3-month period

[1](https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20211216-3" \t "_blank)

.

* However, only 23% of people actually verified the truthfulness of the information or content they encountered

[1](https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20211216-3" \t "_blank)

.

Upload the synthetic graph

Explain the SIR

Show the images

Parameters

Is the graph good?

Hi Michele!

Our group has some questions concerning the NA project that might be worth investigating before the next supervision on Friday. Our network is the one with the email data of a European research institution.

First, we decided to follow Luca's advice from last week and generated a synthetic network with the same number of nodes and edges. You can check the attached CSV file to make sure it is correct.

Second, we came up with conditions for our misinformation spread simulation. It is based on the SIR models however we augmented the recovering conditions according to the feedback from the supervision. The difference is that we have a threshold that does not allow any recovery before the information spreads to at least 5% of the network. Once that point is reached there is a 23% chance for an infected node to recover by itself (the number is selected according to statistics for the EU). Additionally, the recovery rate has a 100% success rate but it transmits only through the neighbors of the recovered nodes. We hope that this model sounds feasible for misinformation spread.

So, given all those modifications, we wanted to ask if we should make a comparison between the original network and the synthetic one. We also have some plots of the simulations for our different case scenarios listed in the project description.

Thank you for your time!

Dimitar

Questions for session 2:

1. We plan to research how the threshold value affects the maximum infection and the time for recovery.
2. Is our recovery rate realistic?
3. Is the difference in our cases significant or not?
4. Should we visualize the synthetic graph?

Diam in org – 9, synt-21

Tasks:

1. Test thresholds, infection rates
2. Visualize the original graph
3. Comparison between cases
4. Histograms for infection in n simulations

Presentation:

1. Show stats for the original network and Visualization

This is the original network that we used for our research. It is a sparse network containing 32.4K nodes and 54.4K edges. For the purpose of our analysis, we have selected the unweighted and undirected version of the network.

Our degree distribution shows that more than 50% of the nodes have a degree < than 10. The network is not scalable.

|  |  |
| --- | --- |
| Density | 0.00010344855751524997 |
| Maximum Degree | 623 |
| Minimum Degree | 1 |
| Average Degree | 3 |
| Assortativity | -0.38162654735878593 |
| Number of Triangles | 146976 |
| Average Number of Triangles | 4 |
| Maximum Number of Triangles | 1615 |
| Average Clustering Coefficient | 0.11268073251806623 |
| Fraction of Closed Triangles | 0.02241862617025909 |
| Maximum K-Core | 22 |
| Lower Bound of Maximum Clique | 12 |

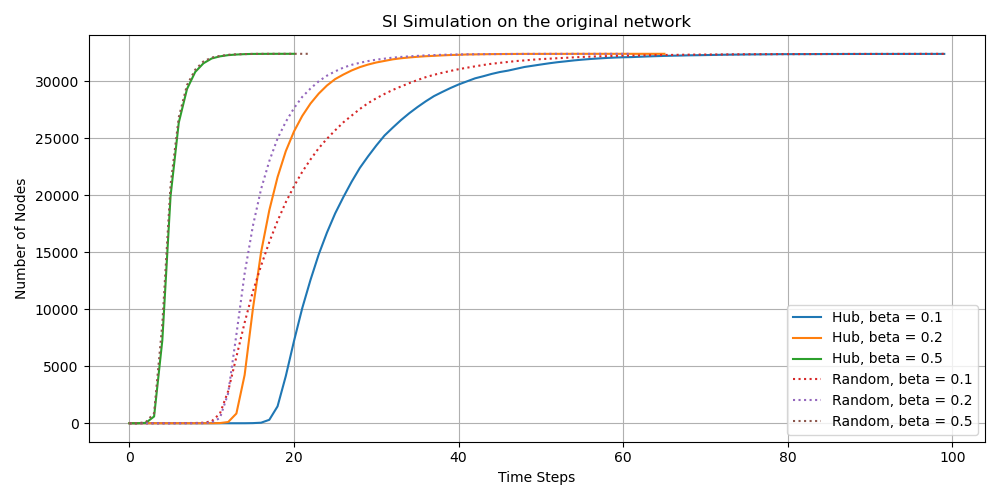
1. RQ

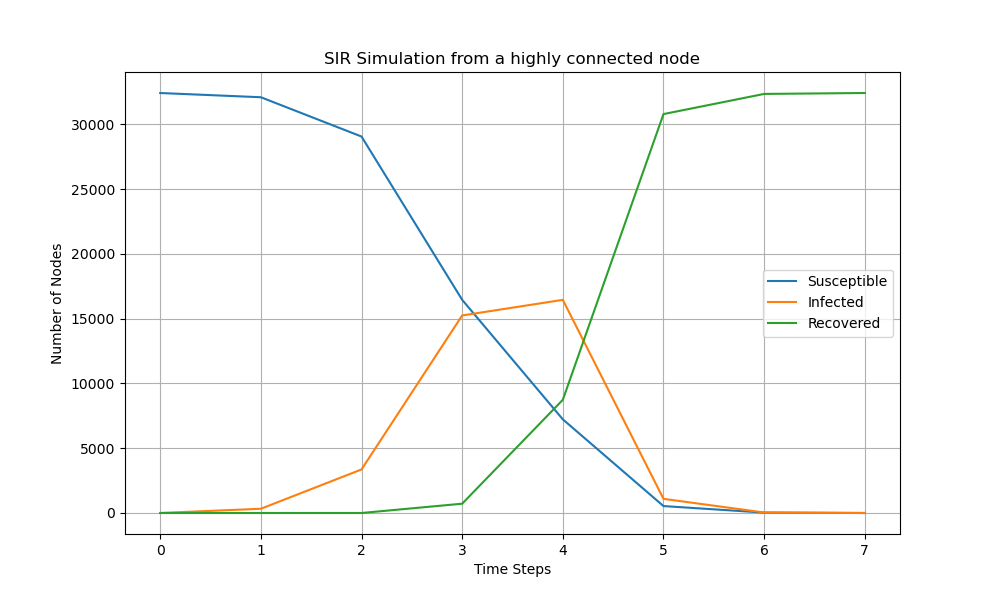
What is the significance of protecting high-degree nodes in misinformation spread? – we expect that hubs are essential for protecting a network.

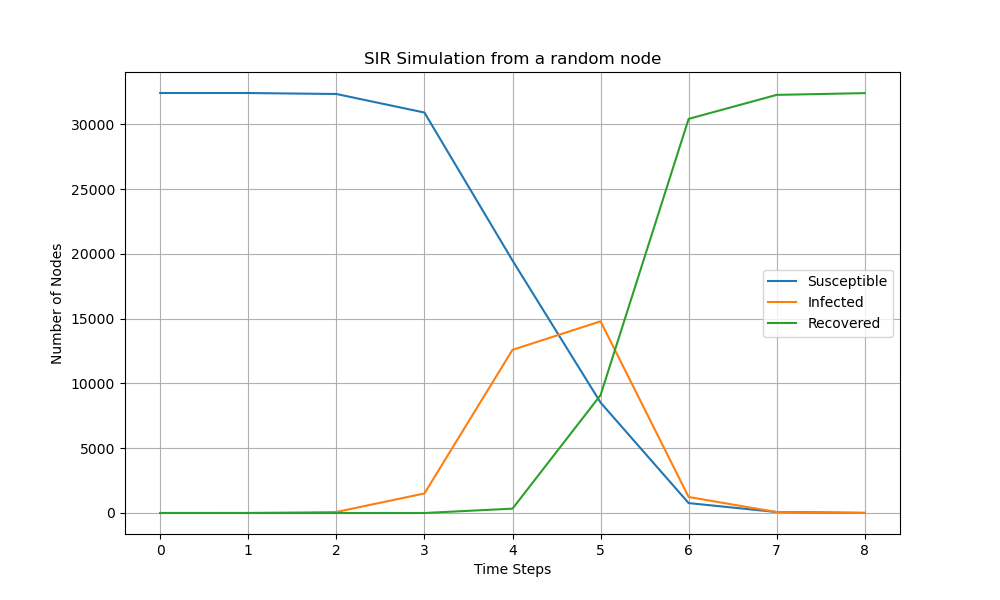
What is the spread behavior in our network? – we know that our network has an unusual structure. How much would this affect our research.

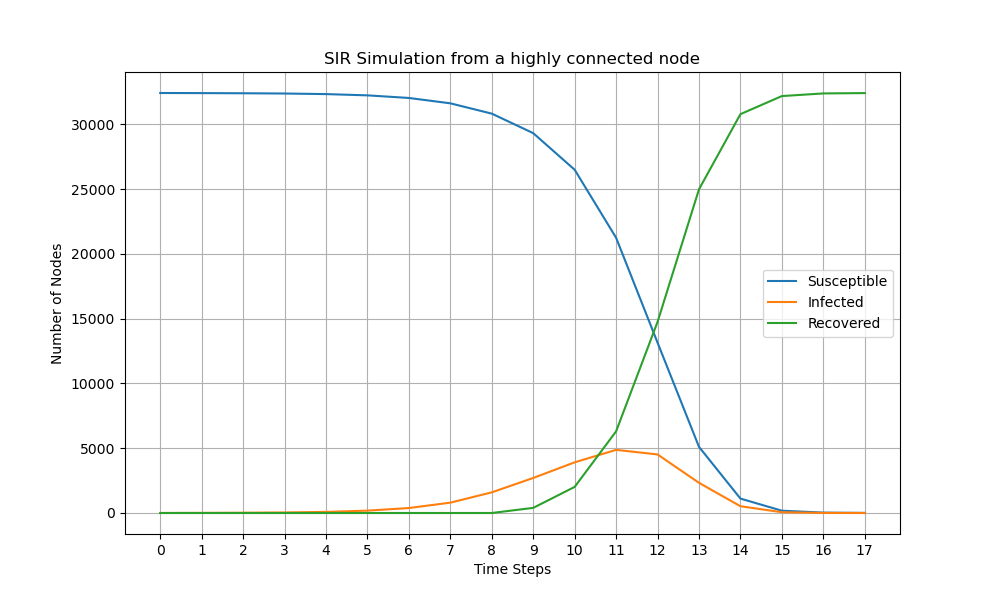
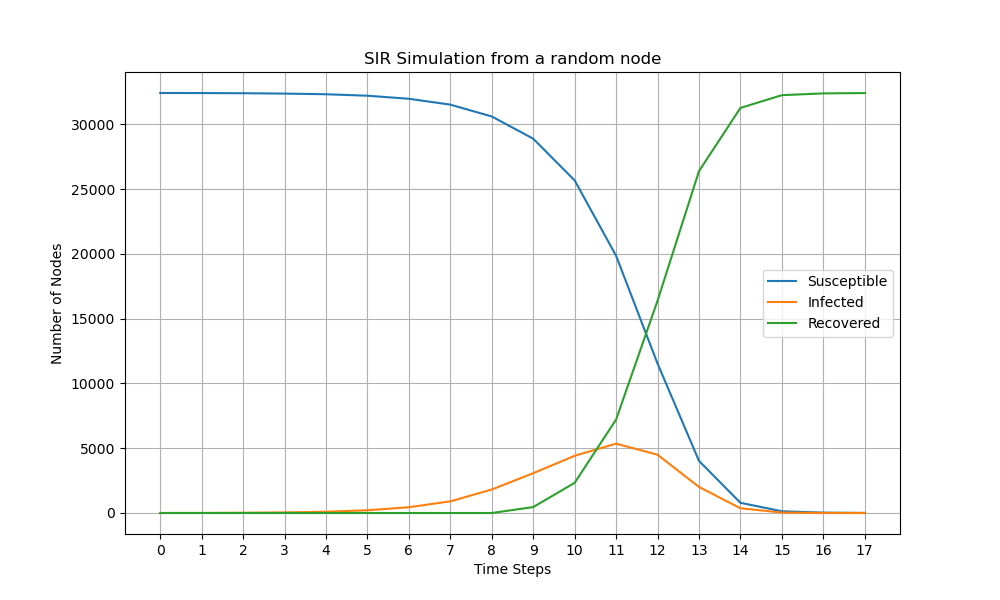
How do we want to define the spread simulations? – Our idea for misinformation spread.

How can we evaluate the results? – Comparison with synthetic graph.

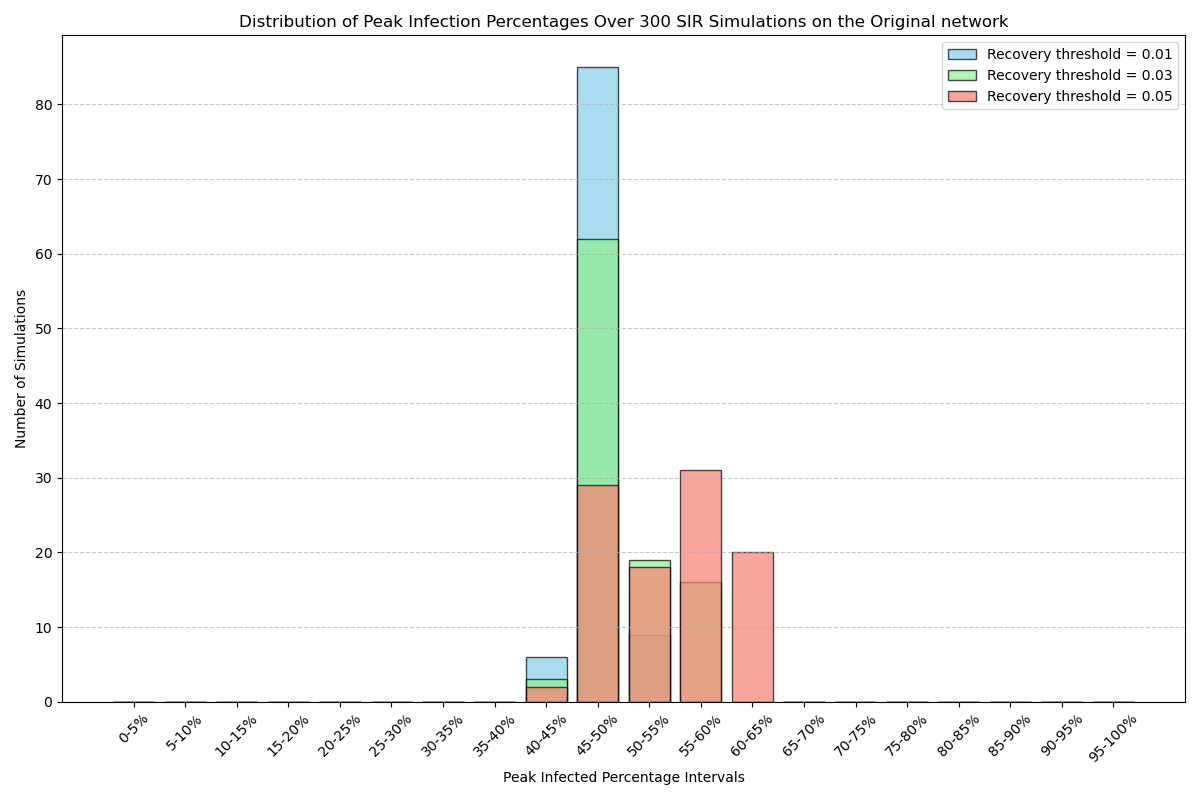
1. Introduce the SI model on the original
   1. First, we examine the spread behavior in the original network.
   2. It does not follow a clear pattern (infection rate is not enough to set the difference).
   3. We chose to continue with beta=0.5 for stable results further in our work.
2. Introduce SIR model
   1. Explain the simulation conditions:
      * Inf rate – 0.5; chosen because of SI
      * initial recovery after spread threshold – 23% taken from EU statistics.
      * spread threshold – 3%; chosen after research with thresholds between 1% and 10%.
      * Recovery probability – 1. We consider that the information can be easily proven. Activates after the threshold. Can turn S to R.
   2. Results
      * Targeting hd node on the original leads to a bigger infection spread. The hypothesis might be true.
      * On synthetic there is no significant difference. H0 requires more investigation.
   3. Significance of threshold – the distribution on the original presents inconsistency between the different cases. There is no pattern. The result of synthetic is clearer and more reasonable.
3. Protected simulations
   1. Explain the simulation conditions:
      * We add 10 protected nodes from the beginning. They never spread the misinformation, but once it reaches them, they activate the recovery process.
   2. Difference compared to the normal SIR
      * Simulations look to take longer but have lower peaks.
      * Anomaly in the original network.
   3. Difference between protection strategies
      * For original: no anomaly; in random, there is a smaller shift to the left; top 10 is significantly longer on average than random
      * For synthetic: in random, there is a smaller shift to the left; very similar results for both protections
4. Conclusion – Tables
5. Further discussion
   1. Page rank is a good alternative because it shows frequently visited nodes. If they are protected the recovery might occur faster, because they have a high probability of misinformation exposure at an early stage.
   2. Betweenness centrality for the determination of protected nodes can provide us with the significance of the structure. That is how we can restrict the misinformation jumping between communities.
   3. SEIR – makes more sense when it is applied to the nature of the problem.
6. Limitations – pretty much written

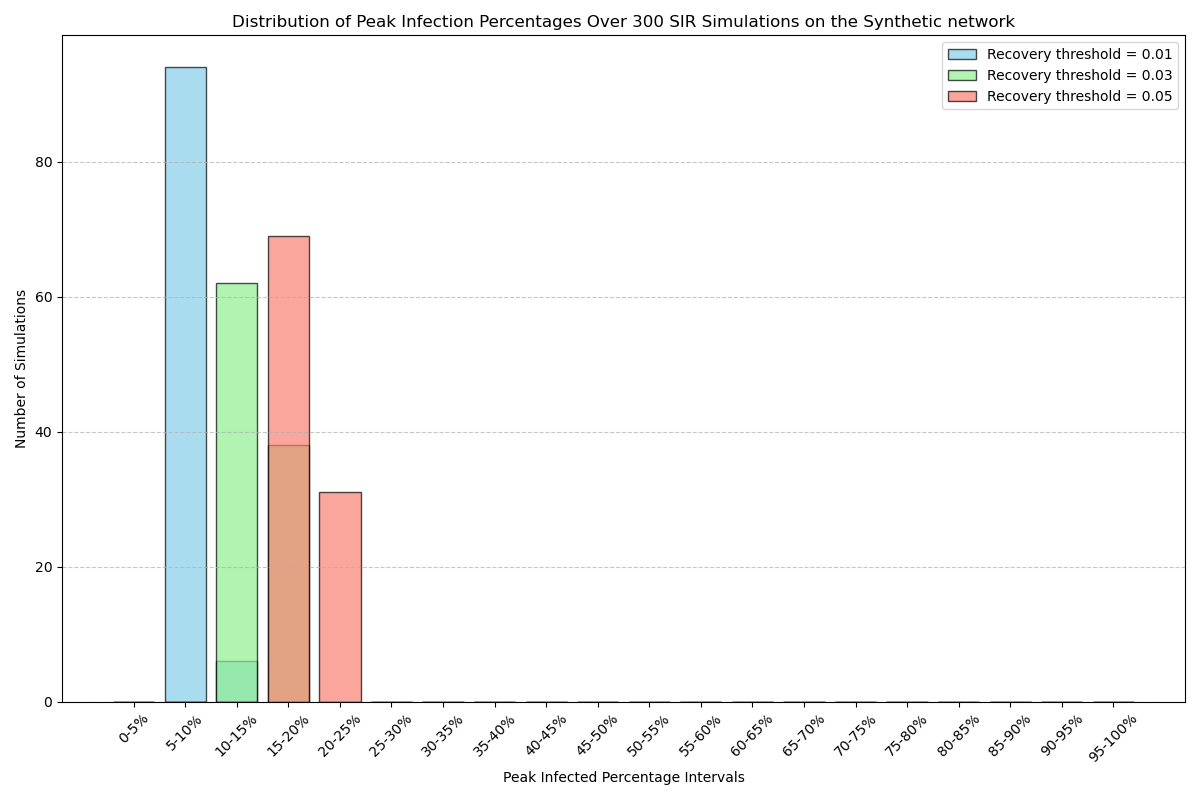






1. Peek Infection distr



1. Conclusion
2. Further discussion
   1. Possible alternative methods

Page rank

Betweenness centrality

* 1. SEIR model

1. Limitations - time

Questions for session 3:

1. Should we make the protected case simulation?
2. How many nodes do we want to protect? Should we discuss how we determine the nodes that should be protected?
3. Should we add more data