Using Complex Networks to understand tourist reviews

Alex **Becheru**, Costin **Badică**, Mihaela **Colhon**
University of Craiova, Romania

Tourist reviews can be useful but often they can be also misleading.

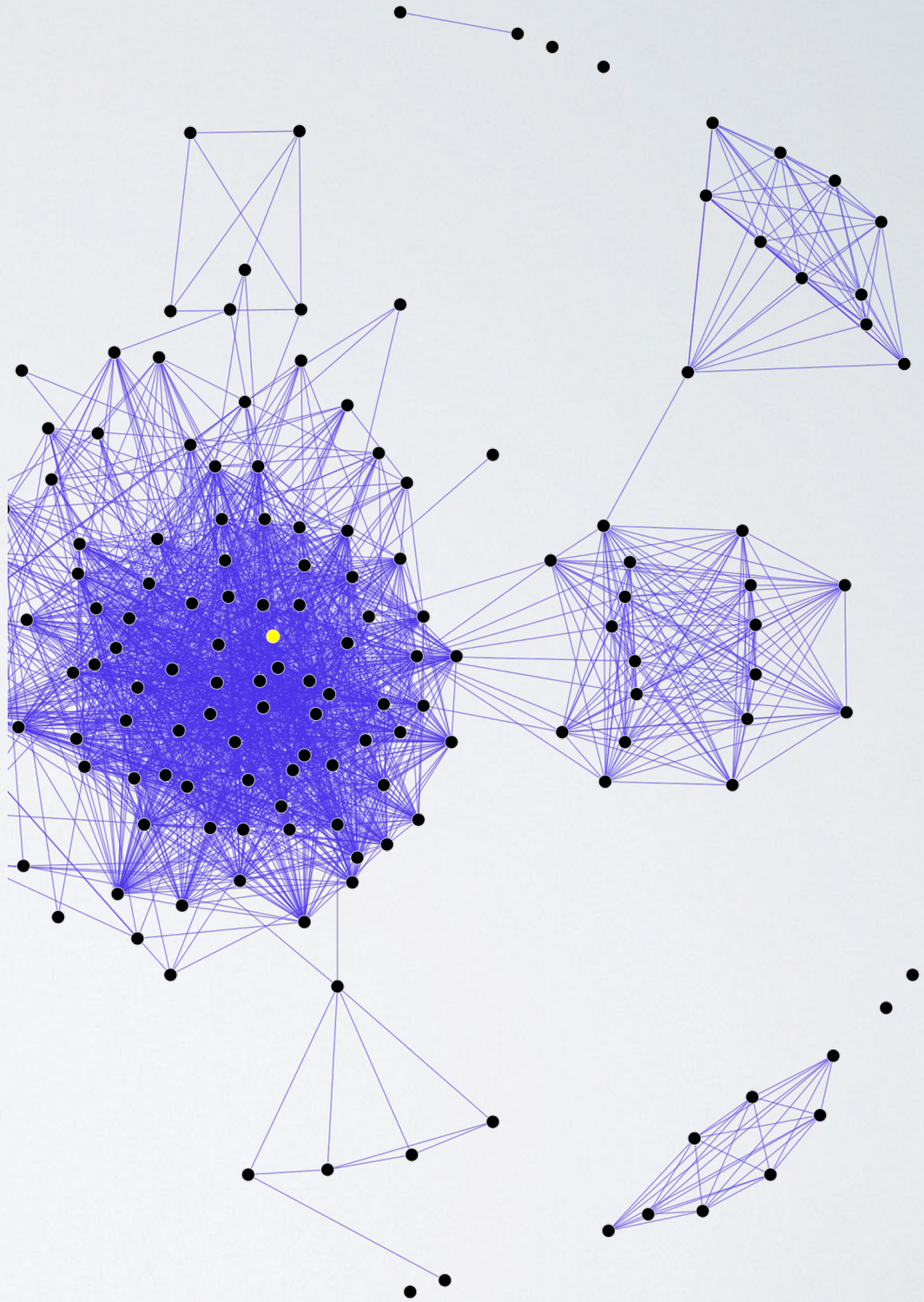
Can we put some order in the tourist reviews?

TOOLS

- Natural Language Processing (**NLP**)
- Complex Networks Analysis (**CNA**)
- Extensible Markup Language (**XML**)
- Programming skills

COMPLEX NETWORKS ANALYSIS (CNA)

- based on **graph theory** and computer science
- investigates **non-trivial features** of graphs that are nor addressed by lattice theory or random graphs
- the complexity comes from **overlapping and interdependent phenomena** present in such networks

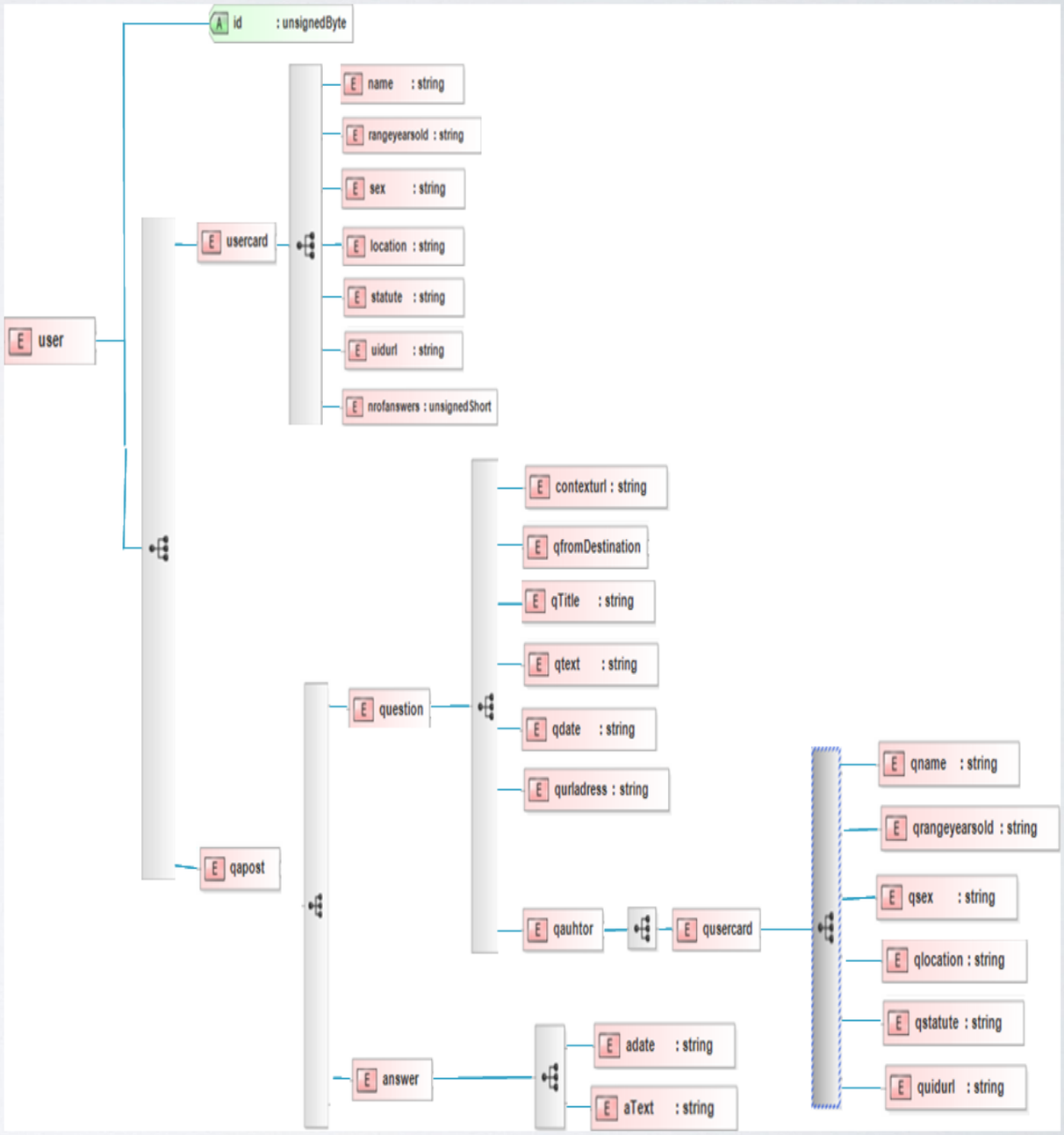


DATA SET

- **amfostacolo.ro** (eng. I was there) web-site
- 45 **countries**, 161 **regions**, 529 **localities**, 1420 **tourist locations**
- 886 sections that do not represent accommodation units, but rather general impressions about a tourist location
- 8017 **users** taken in consideration
- 2527 **comments** considered

DATA ACQUISITION

- developed **web-crawlers**
- parsed the data into 2 XML schemas
  - **user interaction** scheme
  - **reviews scheme**
- gathered **metadata** like:
  - **time & date** of review
  - **location** to which the review addresses
  - **user** that created the review
  - **user given marks**



Initial Question

Answer 1

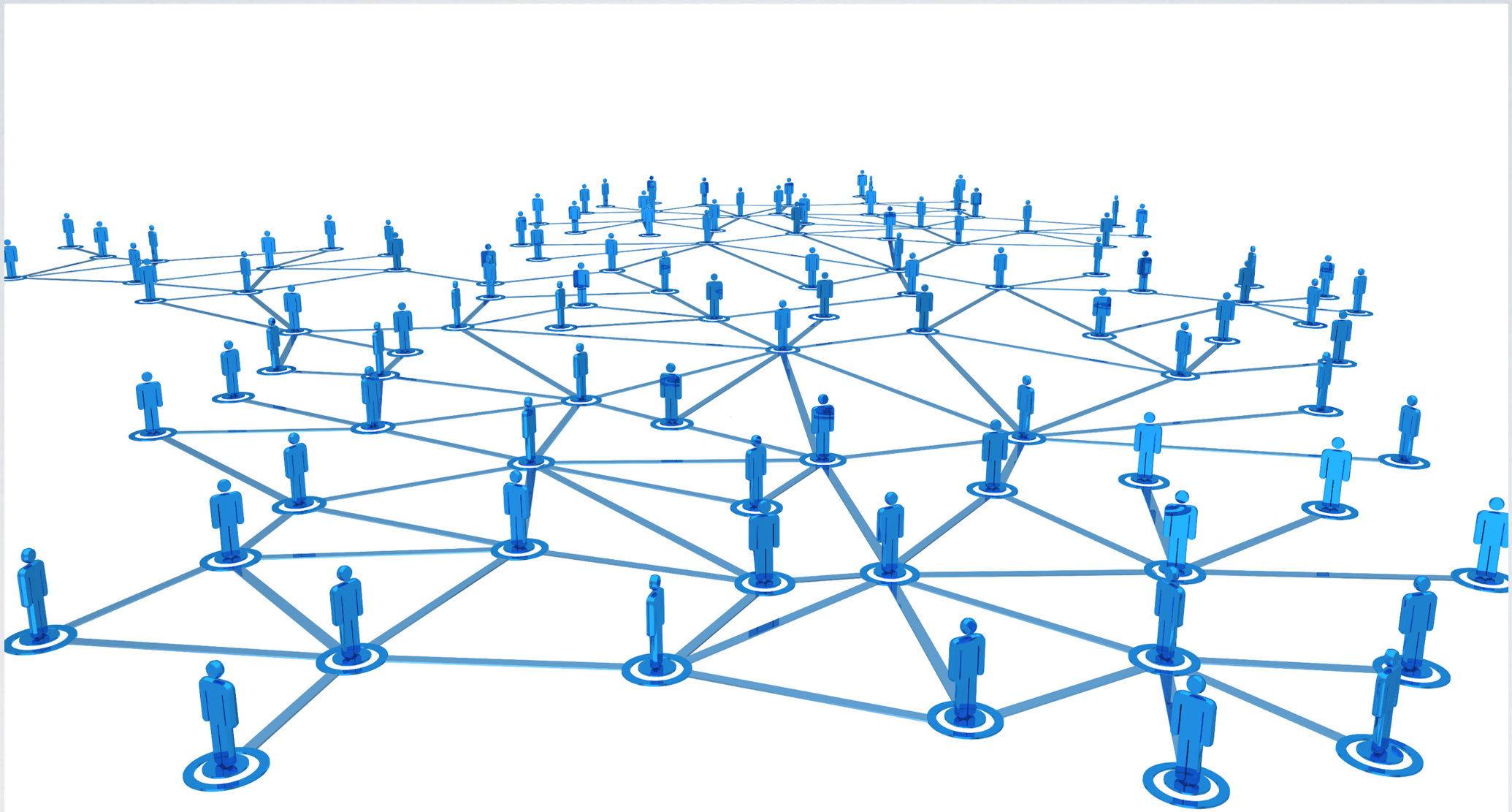
Echo1

Comment on Echo1

Answer 2

Registered **users** are able to **interact** with each other trough:

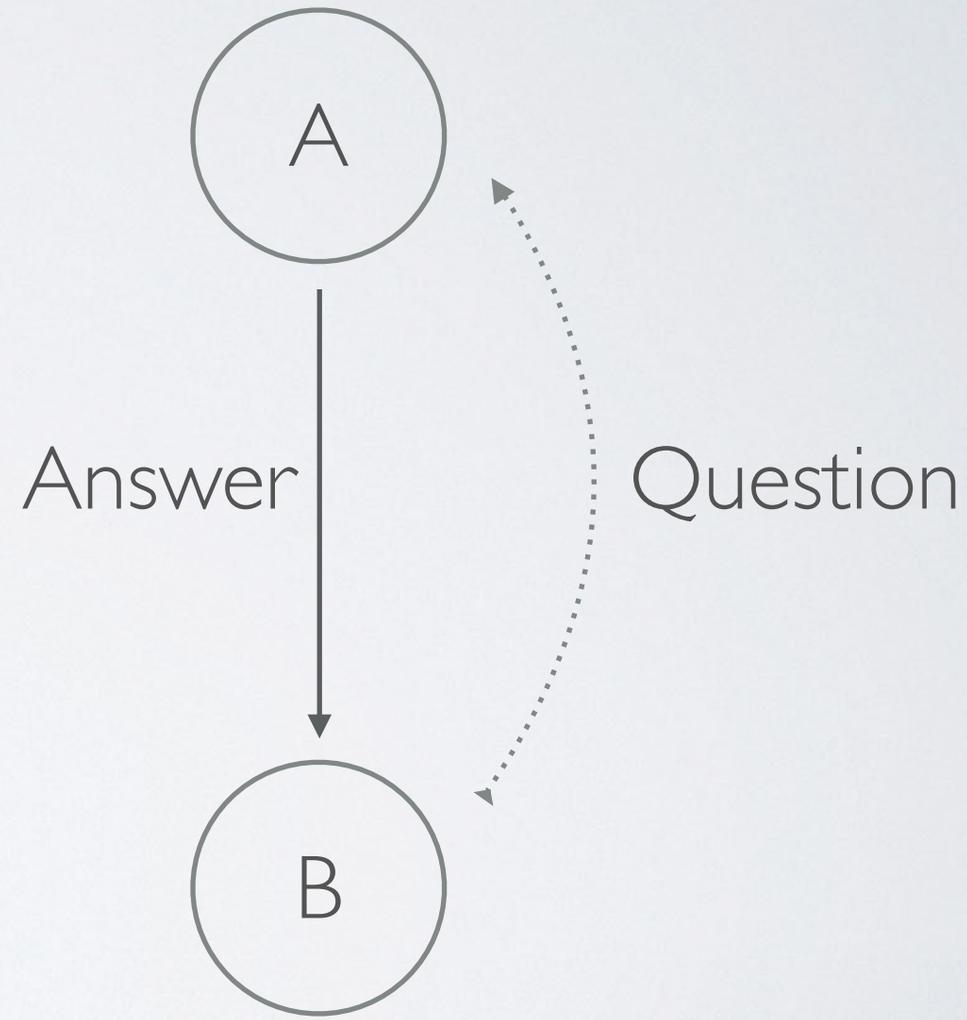
- **echoes**, as well as **answers** to echoes posted in relation to certain reviews or comments
- **asking questions** and giving **answers** to questions about a certain tourism entities.



ANALYSING THE USER INTERACTION NETWORK

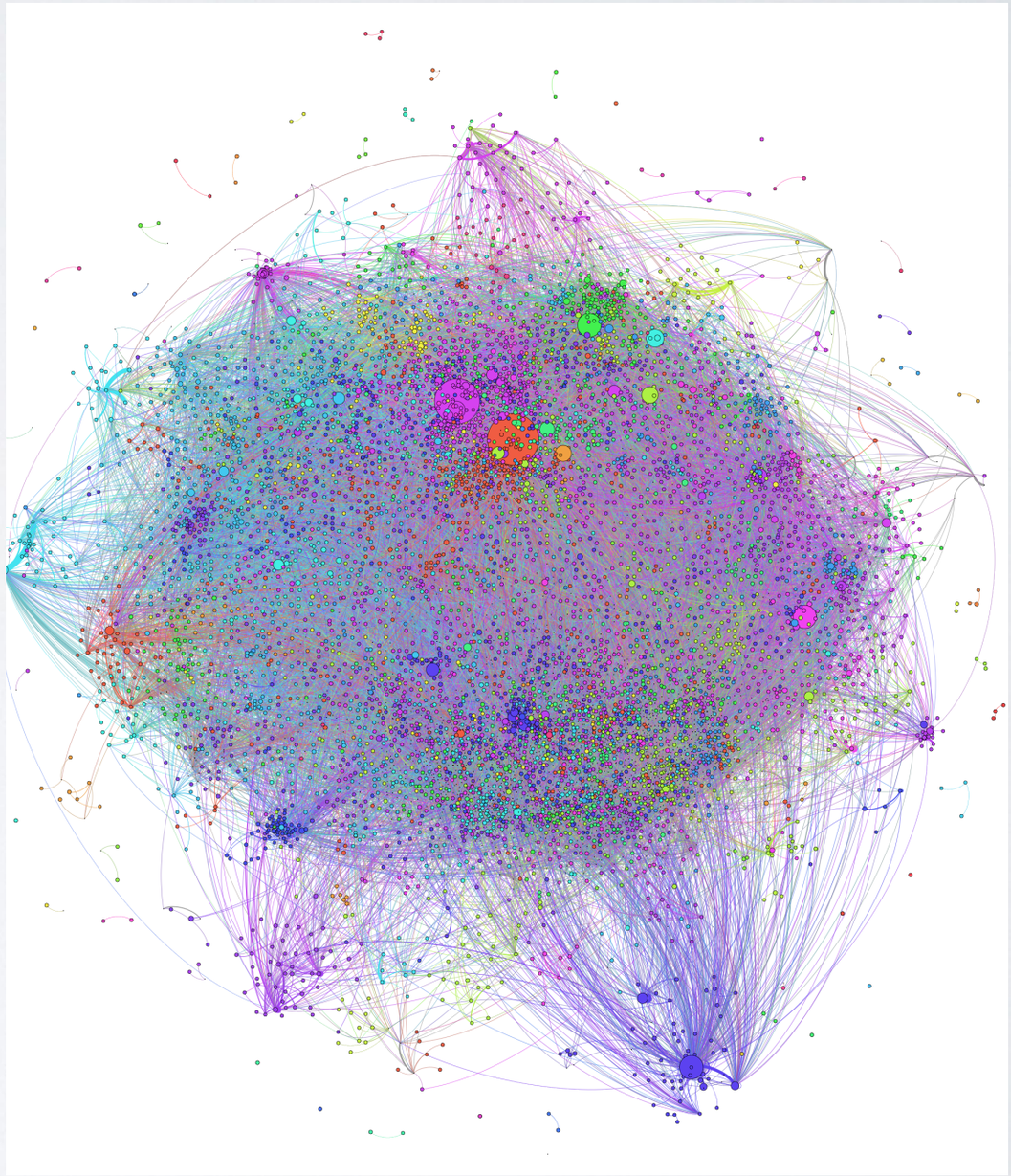
NETWORK CREATION

- a **node** = a **user**
- a **link** between vertices A and B is created when user A responds to a question formulated by B

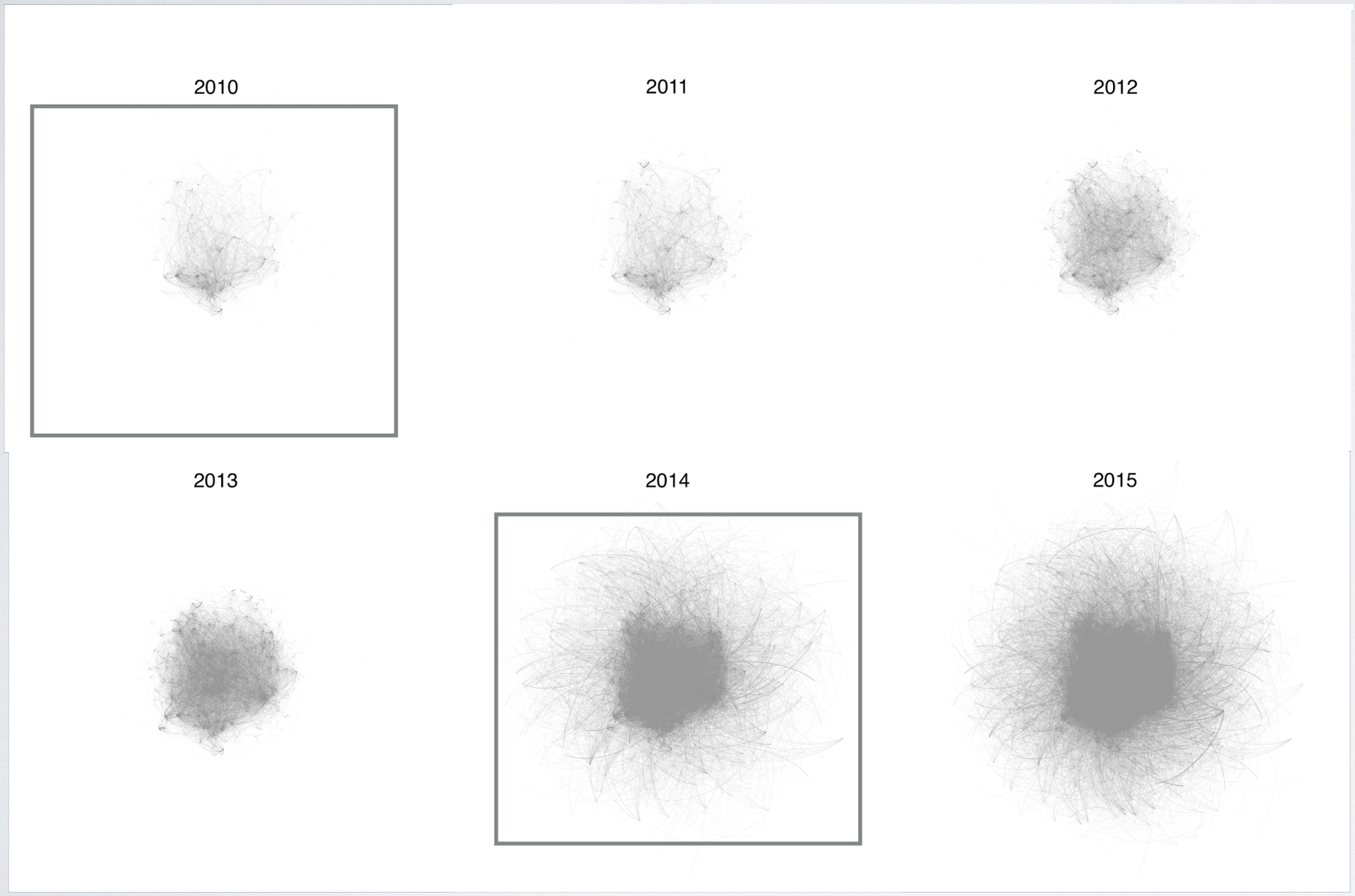


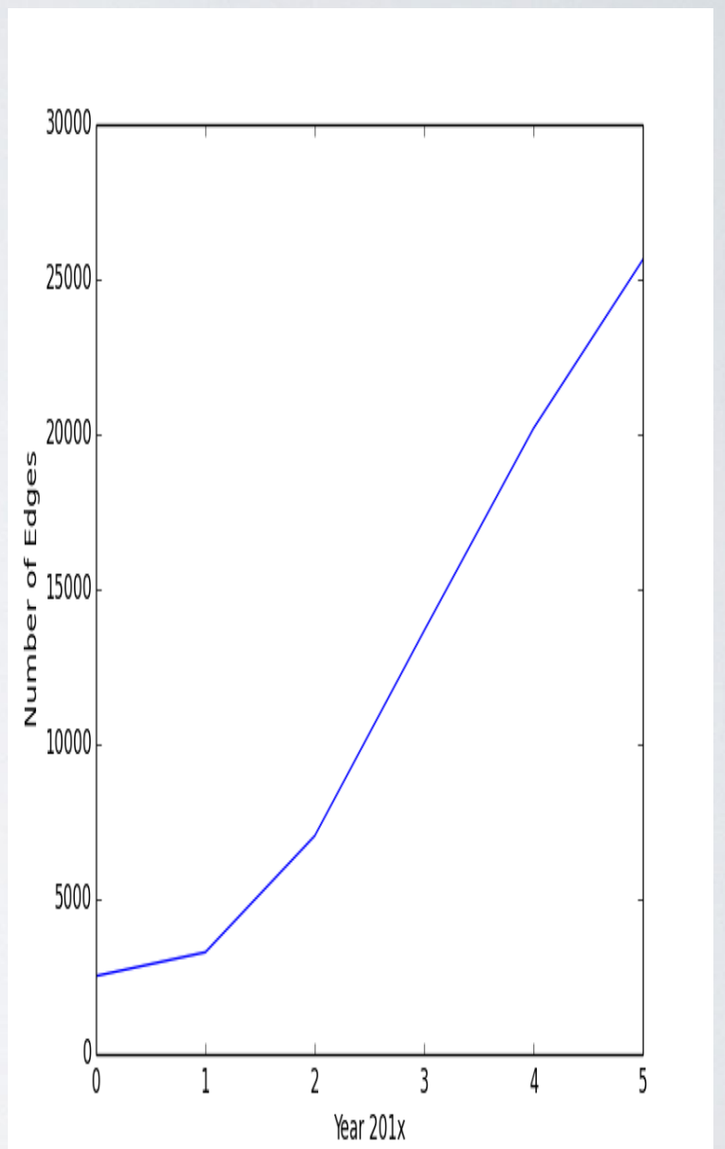
RESULTED NETWORK

- 8017 vertices
- 25666 links



EVOLUTION OF THE NETWORK





NETWORK TYPE

We have proven the network to be of **core-periphery** type, presenting the following characteristics:

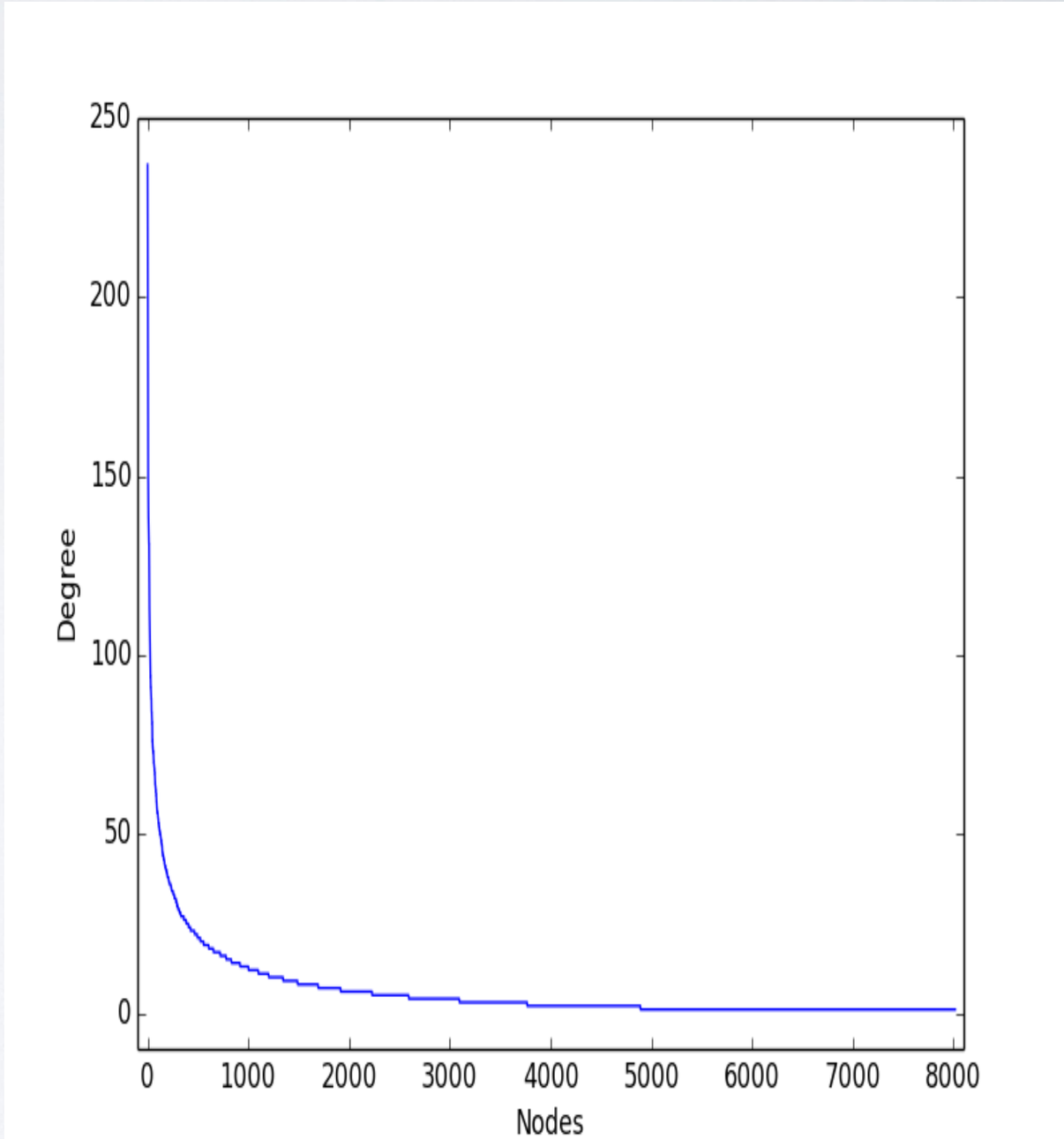
- **high resilience**
- **information exchange** within the network is **fast**
- "**meritocratic**" community, the more you post the more important you become

Metrics/Network	Entire Network	Core
avg. degree	3.2	10.2
diameter	16	9
modularity coefficient	0.48	0.3
avg. path length	0.15	0.68
avg. clusterring coefficient	5	3.2

SOCIAL PHENOMENA

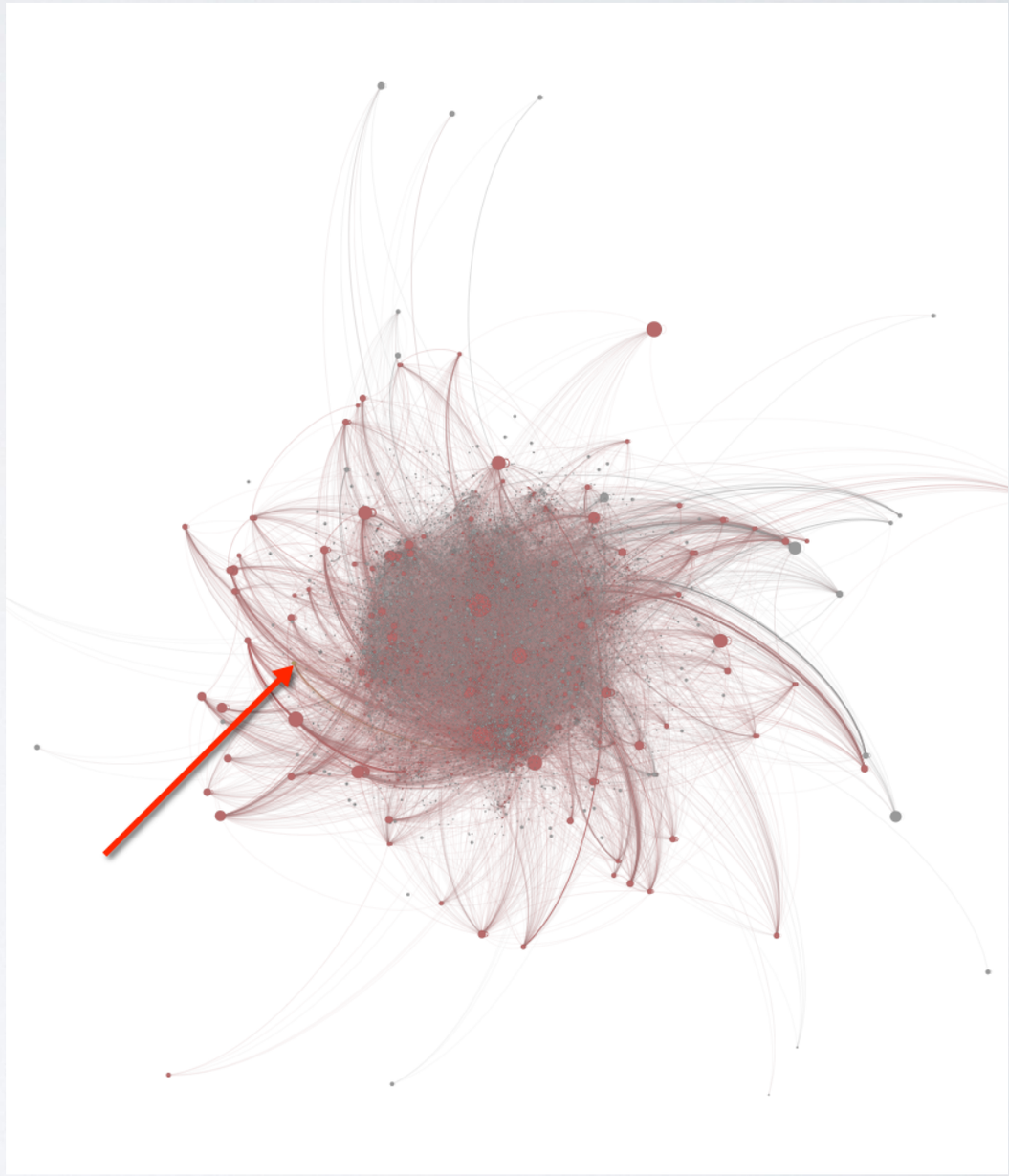
- the presence of hubs betrays the presence of **preferential attachment**

- **small world** phenomenon is also present



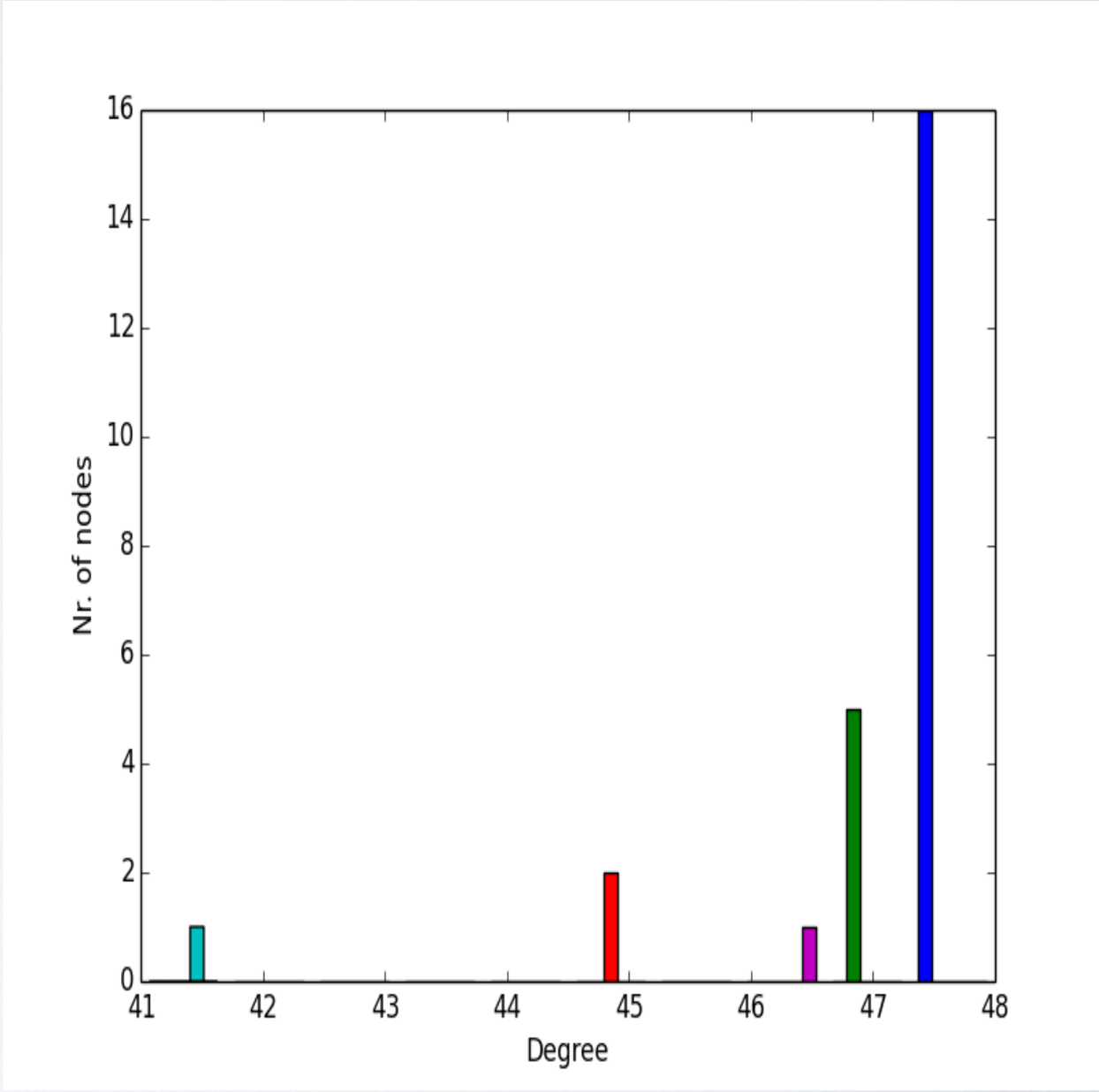
DIFFUSION EXPERIMENT

For the diffusion experiment we chose as start point a **vertice in the periphery** of the graph with out degree 8 (double the average). The diffusion is set to **loose 70%** of its strength at each step. Only the vertices up to neighbour of neighbour can further forward the diffusion, the rest can only receive.



COMMUNITIES

We used **modularity algorithm** to detect network inner communities. We discovered 25 communities that are very well connected among them, the **average inter community degree is 47.2** (48 is the maximum)



GEOGRAPHICAL INTERESTS

The large majority of questions on the web-site refer to a tourism entity, and each **tourism entity** can be pinpointed to a **location**. A location maybe a country, city or a specific address. Thus we were able to **construct a network were vertices represent locations**. A directed link from vertice A to vertice B was constructed if a user from location A answered a question regarding location B.

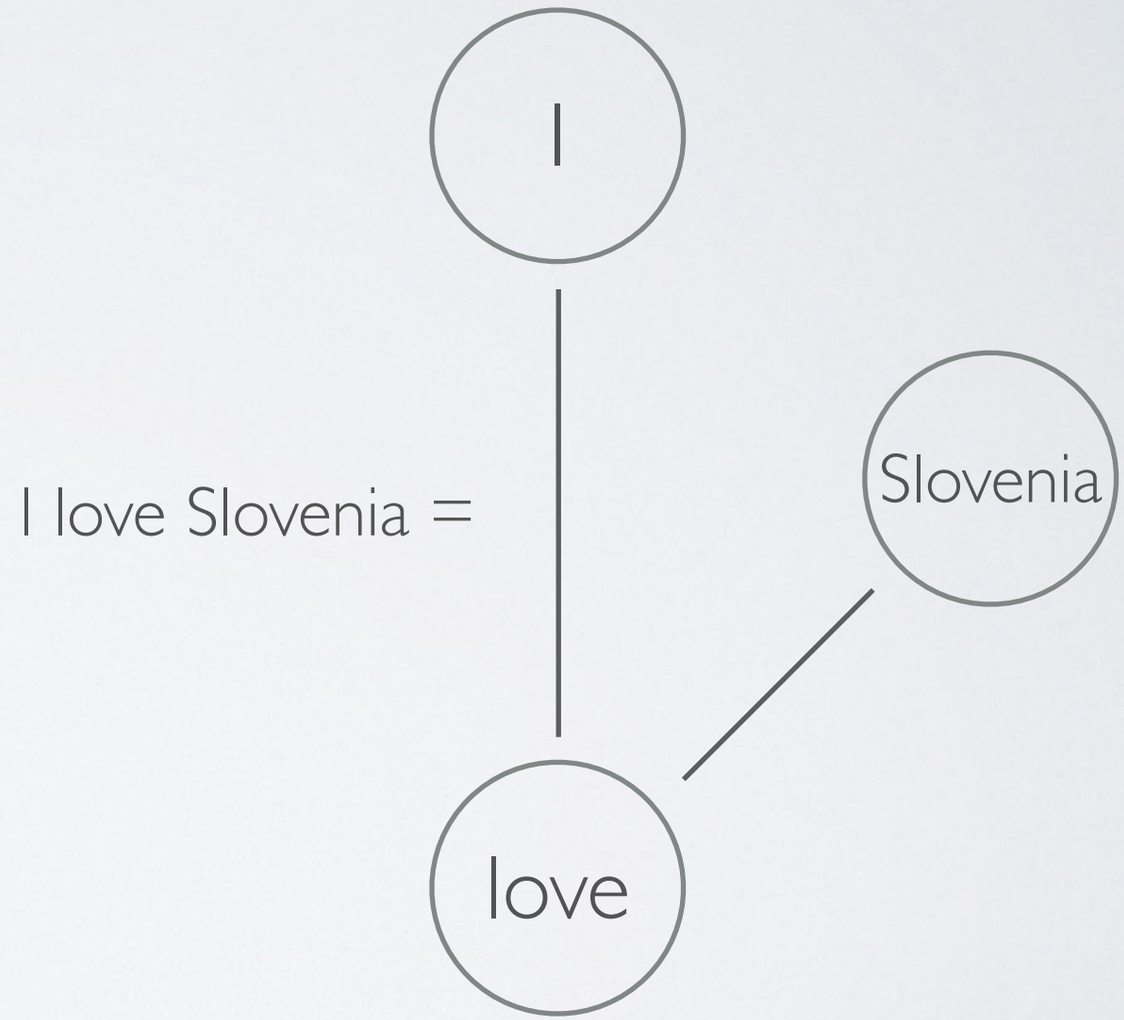
The results are surprisingly accurate at showing the top preferences of the Romanian tourists.

Countries of interest	Magnitude
Greece	100%
Bulgaria	79%
Turkey	45%
Romania	26%
Egipt	19%

REVIEWS ANALYSIS

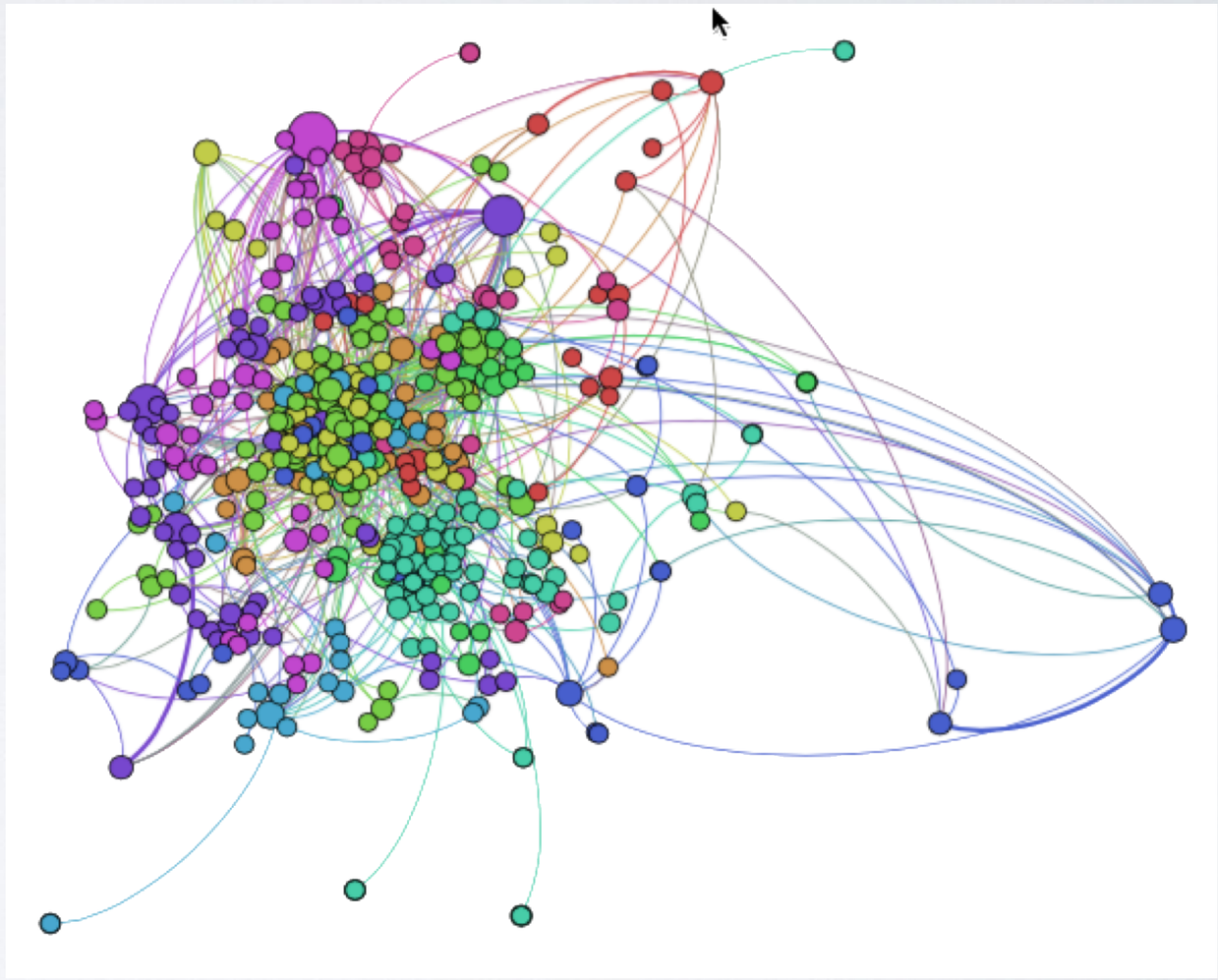
NETWORK CREATION

- a **node** = a **word**
- a **link** between nodes A and B is created if word A comes before/after node B in the same sentence



NETWORK GENERALITIES

- all the networks contain a giant component with 99% of the nodes

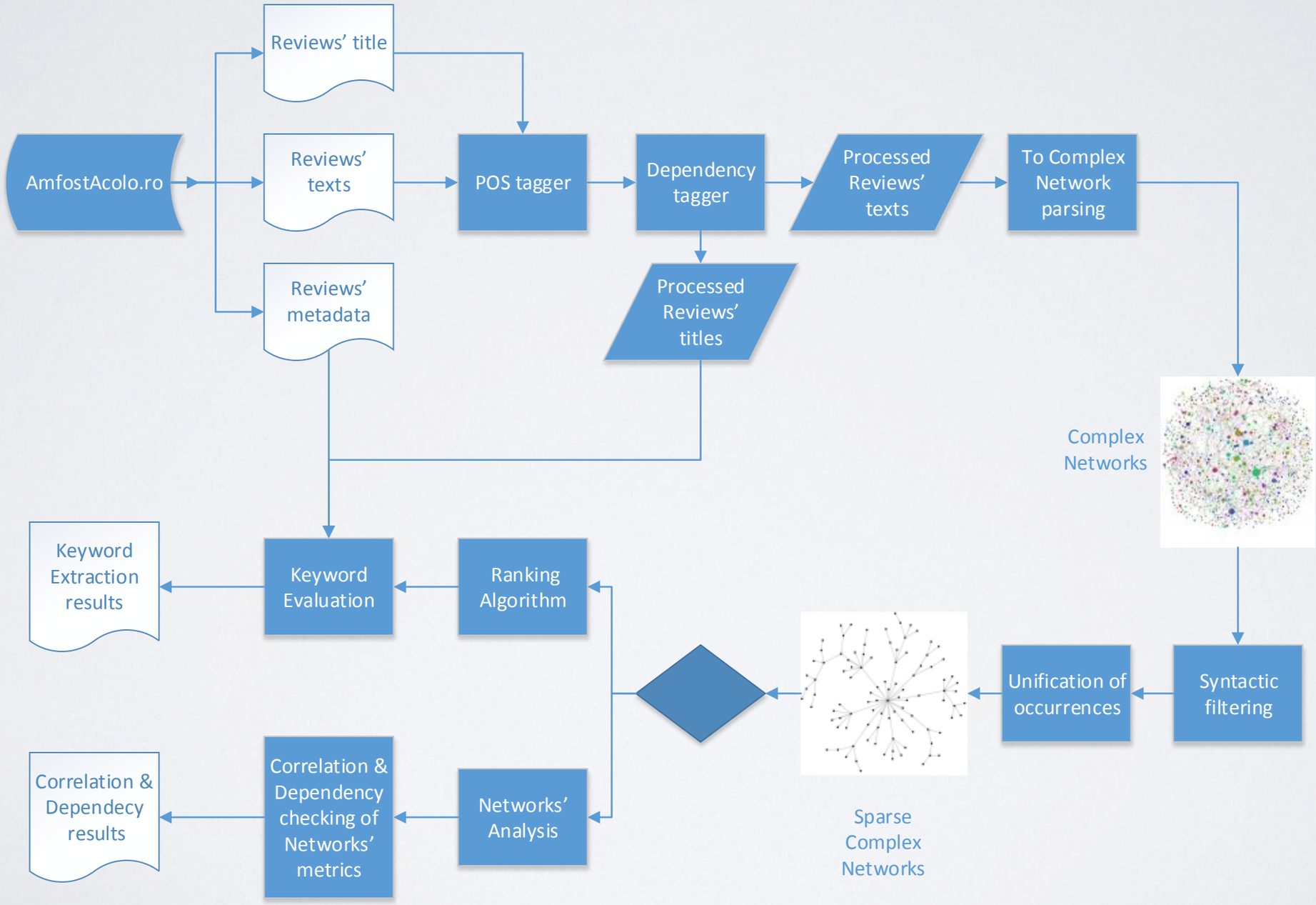


- small world phenomena is present in all the networks

- all the networks are of type core-periphery

- ? do random generated networks have the same characteristics

TEXT SUMMARISATION



RESULTS

Nr.	Method	Title	Text	100%	50%	33%	20%	10%	5%	Nr. reviews
1	degree	NA	NA	0,1199	0,05168	0,0319	0,0233	0,0111	0,0066	2542
2	degree	NA	NA	0,4335	0,1867	0,1154	0,0844	0,0404	0,0240	703
3	degree	NAAdV	NAAdV	0,3671	0,1606	0,0687	0,0989	0,0328	0,0192	723
4	degree	NAAdVM	NAAdV	0,2715	0,1328	0,0837	0,0580	0,0264	0,0139	2184
5	PageRank	NAM	NA	0,2868	0,1400	0,0887	0,0618	0,0271	0,0148	2171
6	PageRank	NAAdVM	NAAdV	0,2717	0,1326	0,0838	0,0579	0,0253	0,0137	2181
7	PageRank	M	NA	0,3688	0,1791	0,1104	0,0748	0,0343	0,0179	2081
8	Tfi Idf	NAM	NA	0,1736	0,0817	0,0593	0,0390	0,0221	0,0129	2107

KEYWORD EXTRACTION RESULTS. THE COLUMNS TITLE AND TEXT INDICATE THE PARTS OF SPEECH THAT WERE KEPT AFTER FILTRATION: N = NOUN, A = ADJECTIVE, AD = ADVERB, V = VERB AND M = LOCATION METADATA. THE FOLLOWING COLUMNS REPRESENT THE NUMBER OF WORDS FROM THE TEXT USED TO MAKE THE COMPARISON. FOR EXAMPLE, THE COLUMN LABELED 20% MEANS THAT ONLY THE TOP 20% OF WORDS FOR THE GIVEN METHOD OF EXTRACTION (INDICATED BY THE ROW LABEL) WERE CONSIDERED. SO, **FOR COLUMN 50% AND ROW 5 WE SHOULD READ THE RESULT AS FOLLOWING: ON AVERAGE, IN 14% OF TEXTS' TITLES WE COULD FIND WORDS FROM THE TOP 50% WORDS RANKED BY PAGERANK WHEN FROM THE TITLE WE CONSIDER ONLY NOUNS, ADJECTIVES AND LOCATION METADATA AND FROM THE TEXTS WE CONSIDER ONLY NOUNS AND ADJECTIVES.** THE NR. REVIEWS COLUMN INDICATES THE NUMBER OF REVIEWS OUT OF 2542 FROM OUR DATA SET ON WHICH THE STATISTICS WERE MADE.

CONCLUSIONS

- Proven that CNA can be useful in this context.
- We determined that either the **PageRank** or **degree** based methods are **better** than **tf-idf** for the task of keyword extraction.
- The work is s**till in progress** so it is early to make any other assumptions.

FUTURE WORK

- more tourism information web-sites need to be added
- a common framework for gathering tourism information needs to be created
- detailed analysis on the communities needs to be conducted
- **A LOT**

Alex Becheru
becheru@gmail.com