

Networks in R - Project 2 - Wiki Norms

Project Description

The network chosen for the project is described at the ICON as follows:


Wikipedia norms (2015)		Informational Web graph 	
<u>Description</u>		<u>Network summary</u>	
The network of Wikipedia pages on editorial norms, in 2015. Nodes are wikipedia entries, and two entries are linked by a directed edge if one hyperlinks to the other. Editorial norms cover content creation, interactions between users, and formal administrative structure among users and admins. Metadata includes page creation time, and norm category.		Edge Type	Hyperlink
		Node Type	Webpage
		Avg Edges	17,235.00
		Avg Nodes	1,976.00
		<u>Graph properties</u>	
		Directed, Unweighted, Metadata	
<u>Citation</u>			
B. Heaberlin and S. DeDeo, "The Evolution of Wikipedia's Norm Network." Future Internet 8(2), 14 (2016)			
Link			
<u>Hosted by</u>			
Hosted by Simon DeDeo			
<u>Network data</u>			
<u>Name</u>	<u>Nodes</u>	<u>Edges</u>	<u>File Size</u> <u>File Type</u> <u>Format</u> <u>Source</u>
Wikinorms-2015	1976	17235	543.00kb txt edgelist Link

image:

The network data is kept at the link

Data Transformation

The network's data is originally provided in its own format in three CVS files: 1. **nodes.csv** - a tab-delimited file containing node properties: id's, names and other attributes. 2. **links.csv** - a comma-delimited file, containing an information about links 3. **topics.csv** - a comma-delimited file, containing the topic distribution for the page in question.

The content of **topics.csv** is ignored for the time being as it contains additional information not required for the project.

The network data was read and transformed to two formats (gml and Pajek) with the **01_transform.R** auxiliary script.

Basic Network Characteristics

The nodes represent web pages, each page describes some Wikipedia social norm. The links are the HTTP reference from one page to another.

It's a directed graph by its nature, no weights assigned to the links.

Property	Value
Vertices, number	1976

Property	Value
Arcs, number	17235
Average degree	17.4
Diameter	9
Acyclic?	FALSE

Interestig to note there are isolated nodes, their percentage is

```
## [1] 4.8
```

Top 20 Nodes

The first 20 nodes with the largest number of degree. The high number of total degree for a node is provided by incoming links. Which makes sense as it's rare to see a page with a lot of outgoing links but it's rather common for a popular page to get a lot of references to it.

## [1] Page Name	Type	In-degree	Out-degree	Total Degree
## [1] -----				
## [1] What_Wikipedia_is_not	Policy	417	79	496
## [1] Neutral_point_of_view	Policy	452	34	486
## [1] Verifiability	Policy	412	35	447
## [1] Identifying_reliable_sources	Guideline	381	39	420
## [1] No_original_research	Policy	290	25	315
## [1] Notability	Guideline	280	35	315
## [1] Consensus	Policy	274	32	306
## [1] Assume_good_faith	Guideline	275	29	304
## [1] Administrators	Policy	238	34	272
## [1] Biographies_of_living_persons	Policy	223	45	268
## [1] Policies_and_guidelines	Policy	229	34	263
## [1] Civility	Policy	217	28	245
## [1] Criteria_for_speedy_deletion	Policy	168	73	241
## [1] Blocking_policy	Policy	178	49	227
## [1] Vandalism	Policy	165	59	224
## [1] Dispute_resolution	Policy	170	44	214
## [1] Edit_warring	Policy	185	29	214
## [1] Citing_sources	Guideline	170	38	208
## [1] No_personal_attacks	Policy	165	32	197
## [1] Deletion_policy	Policy	123	72	195

Components

Density

The proportion of present edges from all possible edges in the network.

```
## [1] 0.004416287
```

The nodes along the first found path of the diameter distance.

```
V(net)$PageName[get.diameter(net)]
```

```
## [1] "Added_or_removed_characters" "Bypass_your_cache"
## [3] "Purge"                      "Categorization"
## [5] "User_categories"            "WikiFauna"
## [7] "Wikipediaholic"            "The_Wikipedian's_Prayer"
## [9] "Confessed_wikipediaholics"  "Wikipediaholics_in_denial"
```

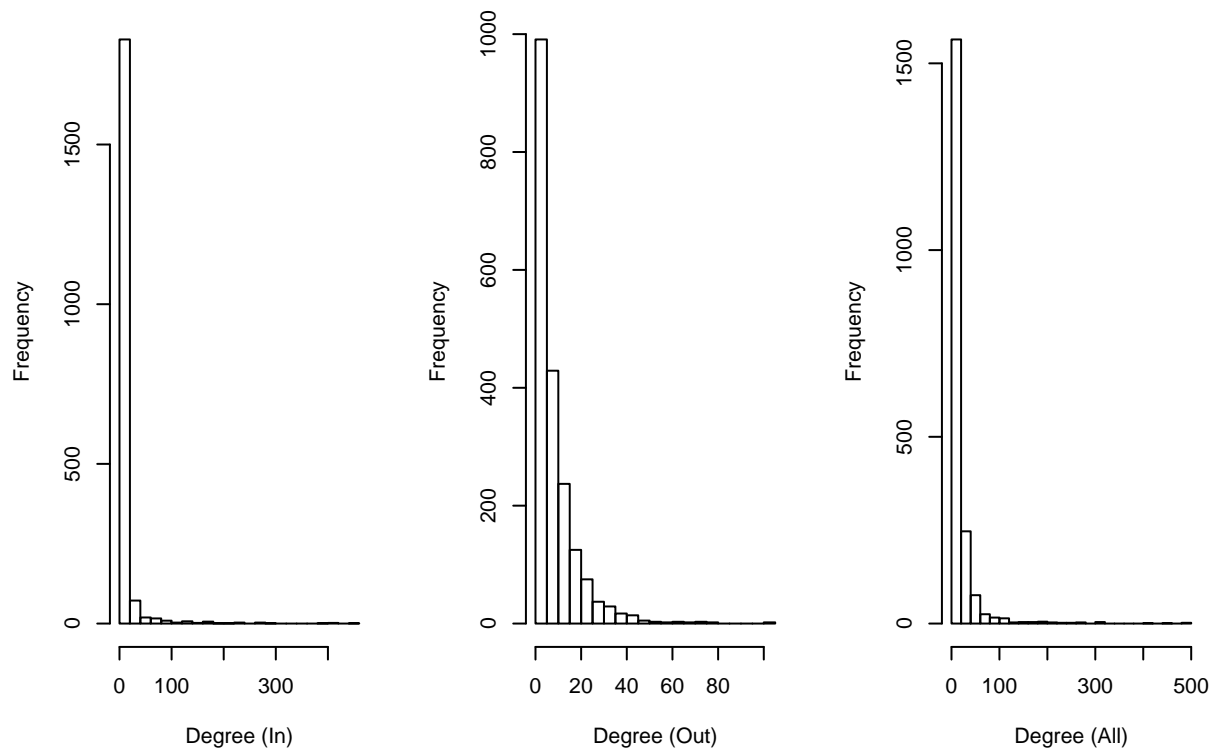
Degree

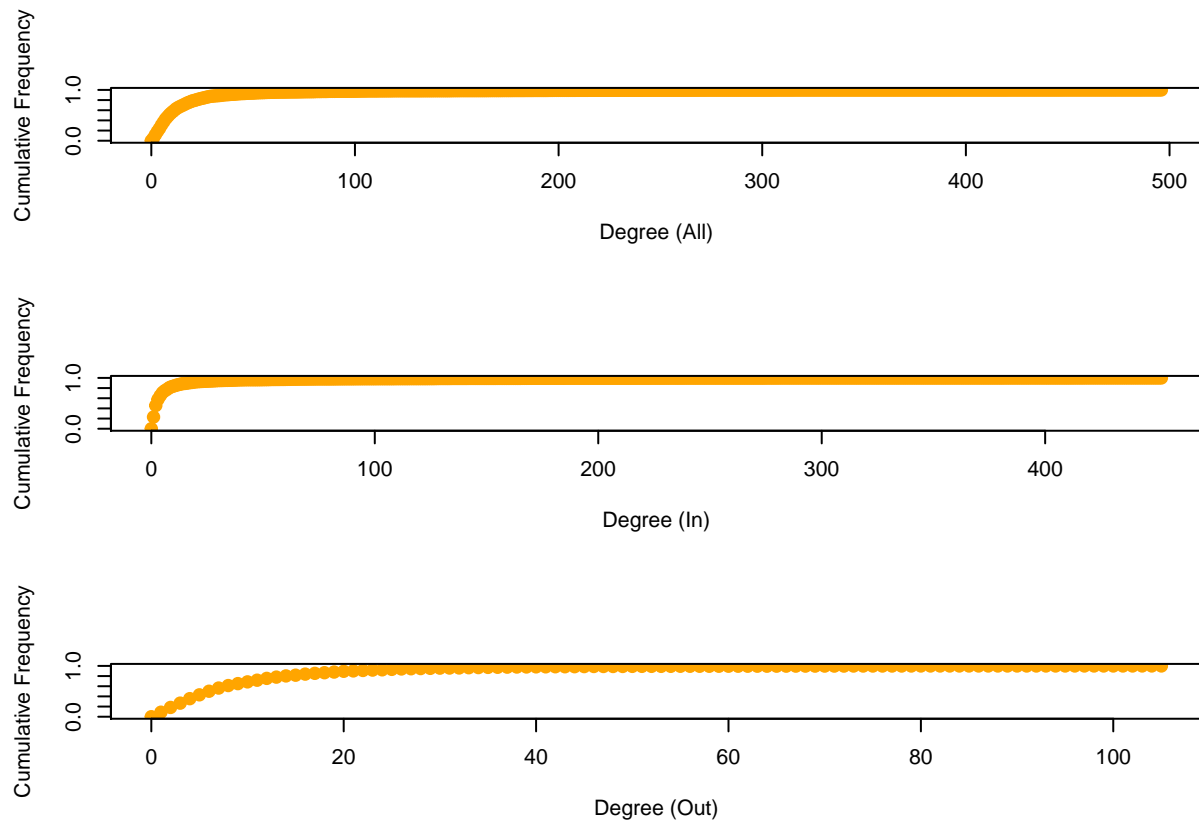
The top nodes in terms of degree were listed previously.

The degree distribution (in-degree, out-degree, total)

```
# Degree
net_deg_in  <- degree(net, mode = "in")
net_deg_out <- degree(net, mode = "out")
net_deg_all <- degree(net, mode = "all")

par(mfrow = c(1,3))
hist(net_deg_in, breaks = 20, freq = T, main = "", xlab = "Degree (In)")
hist(net_deg_out, breaks = 20, freq = T, main = "", xlab = "Degree (Out)")
hist(net_deg_all, breaks = 20, freq = T, main = "", xlab = "Degree (All)")
```





Nodes

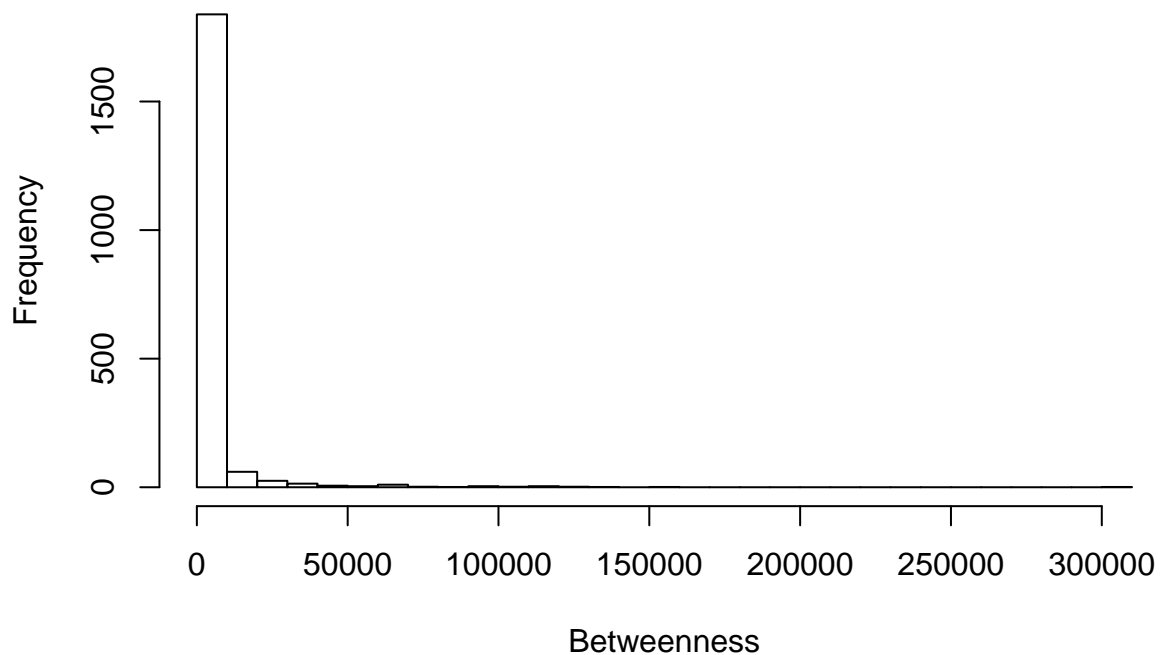
The nodes, as previously said, are pages describing social norms of Wikipedia. The measurement of importance of nodes depends on a research question. For example, if we are interested in mostly referenced pages or mostly edited pages.

From the editorial point of view that's probably the mostly linked page is mostly important i.e. the importance is measured by the number of in-degree. And, the largest nodes in terms of that were previously listed.

->

Betweenness

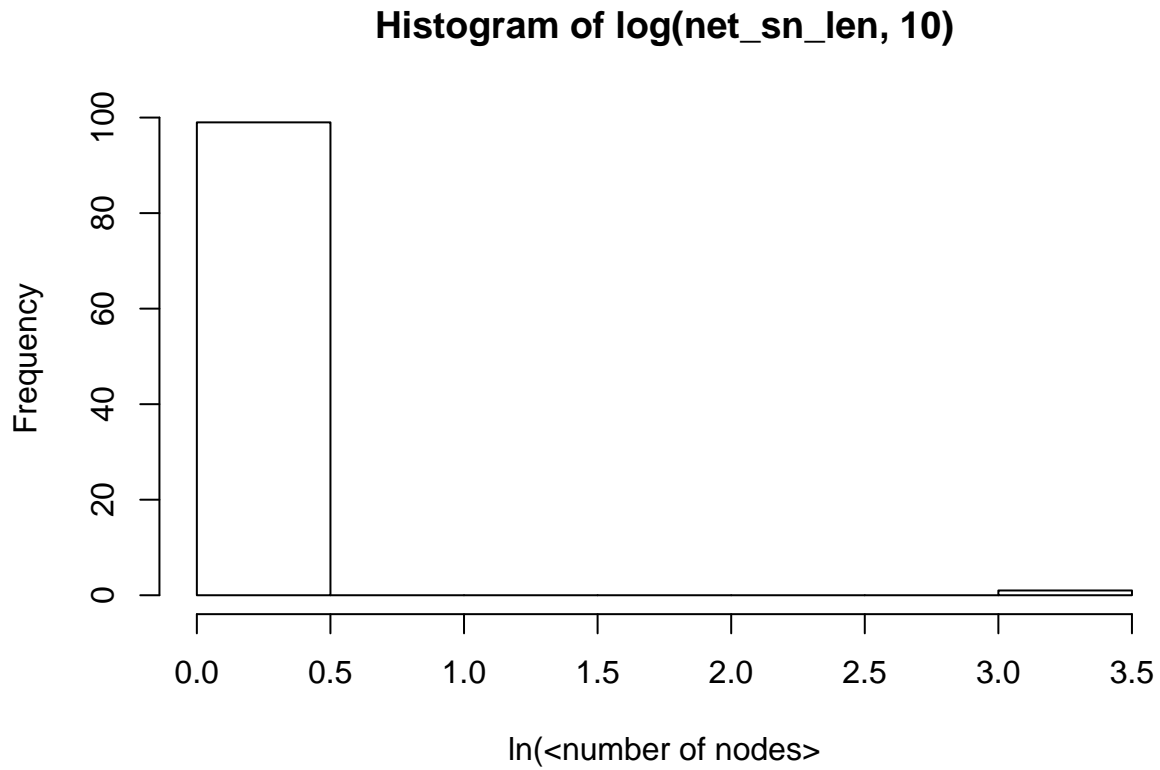
```
net_bt = betweenness(net, v = V(net), directed = TRUE)
hist(net_bt, breaks = 30, freq = T, main = "", xlab = "Betweenness")
```



```
# net_eb <- cluster_edge_betweenness(net, directed = TRUE)
# plot_dendrogram(net_eb)
```

Subnetworks

In trying to determine subnetworks we find there are a hundred of one node networks (isolated nodes) and one large subnetwork of 1872 nodes. That's shown on the histogram below where x is ln of the size of each of those networks



Communities

TBC

Visualization

Visualization was made in Gephi with manual configuration. An effort of doing the same with igraph provides suboptimal results of nodes and edges overlapping each other.

The size of nodes reflects the number of degrees for that node. The color of node reflects which community it belong to.

The visualization demonstrates the network is centered around several large nodes.

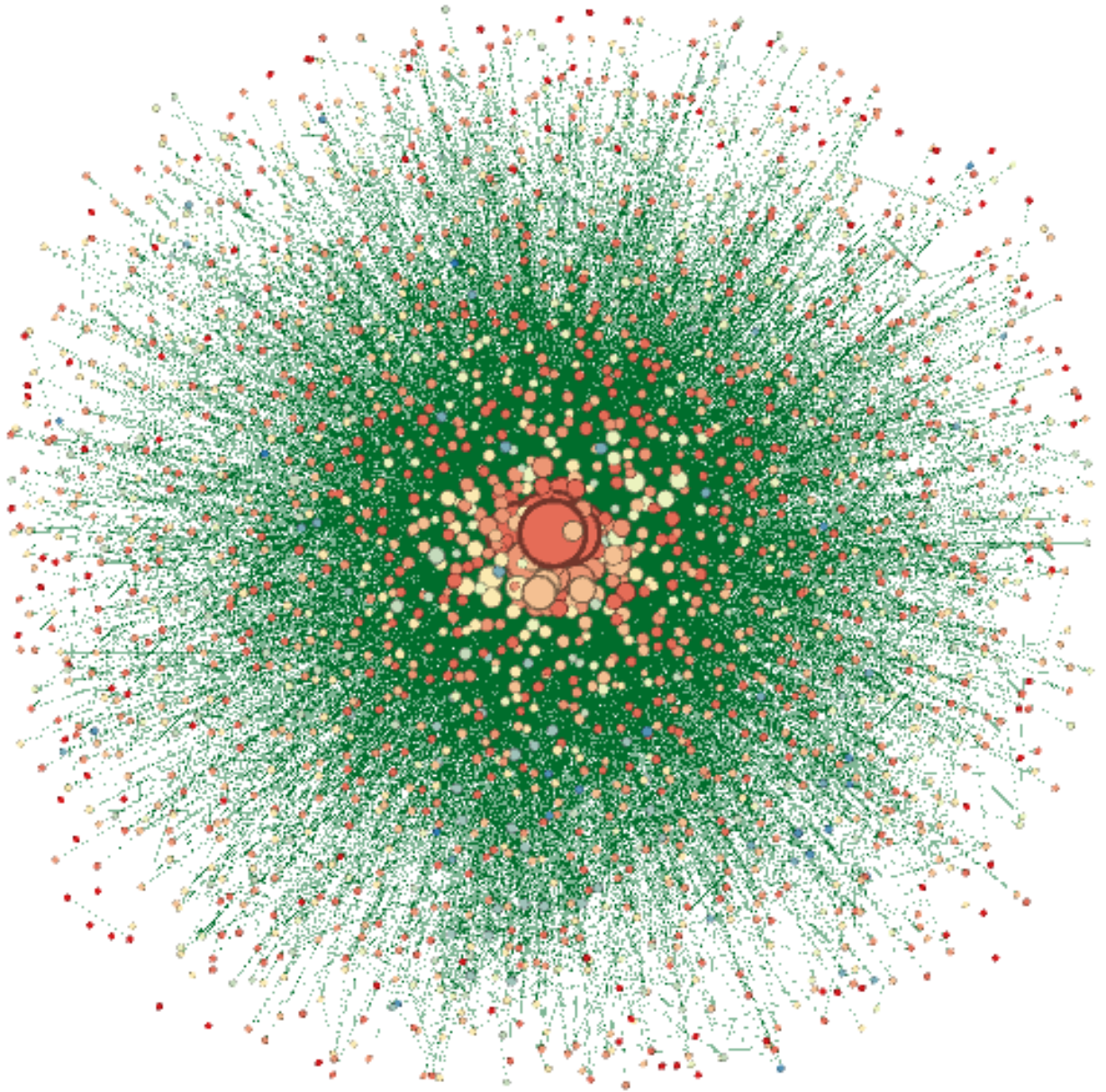


image:

The visualization also saved as `wiki_norms_vis.pdf`

Ideas to Improve Report

1. Take into account the network was already analyzed for communities. Draw some conclusions from it.
2. Provide statistics on betweenness, closeness.
3. Dendrogram? Will it work for such a large network?
4. Try the hierarchical clustering.
5. Provide more interpretation of the network.

Appendix A Technical Details of Report

This version of the report was built with:

```
devtools::session_info()
```

```
## Session info -----
##   setting  value
##   version  R version 3.4.3 (2017-11-30)
##   system   x86_64, mingw32
##   ui       RTerm
##   language en
##   collate   Russian_Russia.1251
##   tz        Europe/Moscow
##   date      2018-01-07

## Packages -----
##   package * version date          source
##   backports 1.1.2 2017-12-13 CRAN (R 3.4.3)
##   base      * 3.4.3 2017-12-06 local
##   compiler  3.4.3 2017-12-06 local
##   datasets  * 3.4.3 2017-12-06 local
##   devtools  1.13.4 2017-11-09 CRAN (R 3.4.3)
##   digest     0.6.13 2017-12-14 CRAN (R 3.4.3)
##   evaluate   0.10.1 2017-06-24 CRAN (R 3.4.3)
##   graphics  * 3.4.3 2017-12-06 local
##   grDevices * 3.4.3 2017-12-06 local
##   htmltools  0.3.6 2017-04-28 CRAN (R 3.4.3)
##   igraph     * 1.1.2 2017-07-21 CRAN (R 3.4.3)
##   knitr      1.18   2017-12-27 CRAN (R 3.4.3)
##   magrittr   1.5    2014-11-22 CRAN (R 3.4.3)
##   memoise    1.1.0  2017-04-21 CRAN (R 3.4.3)
##   methods   * 3.4.3 2017-12-06 local
##   pkgconfig  2.0.1  2017-03-21 CRAN (R 3.4.3)
##   Rcpp       0.12.14 2017-11-23 CRAN (R 3.4.3)
##   rmarkdown  1.8     2017-11-17 CRAN (R 3.4.3)
##   rprojroot  1.3-1   2017-12-18 CRAN (R 3.4.3)
##   stats     * 3.4.3 2017-12-06 local
##   stringi    1.1.6  2017-11-17 CRAN (R 3.4.2)
##   stringr    1.2.0  2017-02-18 CRAN (R 3.4.3)
##   tools      3.4.3  2017-12-06 local
##   utils     * 3.4.3 2017-12-06 local
##   withr      2.1.1  2017-12-19 CRAN (R 3.4.3)
##   yaml       2.1.16 2017-12-12 CRAN (R 3.4.3)
```