AUEB
MANAGEMENT
SCIENCE AND
TECHNOLOGY

FEBRUARY
2022

SOCIAL NETWORKS ANALYSIS

Co - Tag Network on Crypto Community with Tumblr

ΗΛΙΑΔΗΣ ΒΙΚΤΩΡΑΣ 8180026

Contents

Section 1 : Introduction

One of the essential features of many web 2.0 services is social tagging , it can be fun ,accurate, flexible, interesting and most importantly utilized for marketing and information organizing purposes. It is how internet users categorize their media and annotate web resources according to their own needs and understanding. For the purposes of this assignment, we are going to extract a co-tag network from the social blogging platform Tumblr , with the word crypto as its central node that every single other node is going to be directly connected to. In particular the nodes are going to be the various tags we found to co-exist with the central one in the 1946 blogs we extracted , and an edge is going to represent the co-existence of two tags in the same blog . The aim of this report is to examine social tags and their co-occurrence relationships within the structure of an information network, using the social network analysis methods to analyze the basic characteristics of said network so can get a better understanding of how information is linked and spread out in the targeted social community of cryptocurrencies . The final and desired goal would be to get a better grasp of what is trending in the very volatile world of cryptocurrencies , discover connections to topics we hadn't realized exist and perhaps understand where the crypto frenzy is headed.

Section 2 : Data Extraction and Cleaning

Data Extraction

To extract data i initially tried the Twitter Exporter from Gephi but my key was cancelled twice , and then the first level of access is no longer enough to use the plugin , my application for elevated access is still pending , then the Enron dataset caught my attention, but it didn't meet the assignment's conditions of extracting the data. So finally I headed to Bernhard Reiner's Tumblr Tool and used as a central tag-node the word crypto to get a co-tag network from the last 1946 blog's from Tumblr that included my central tag , and download it all as a GDF file .

Data Cleaning

Transferring the data to Gephi , a small percentage of labels got corrupted probably by wrong encoding , some others where incomprehensible . In both cases i removed the node all together.

Ex.

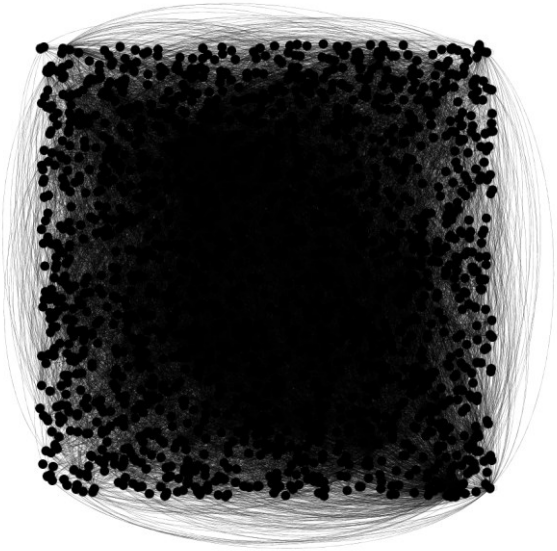
Id	Label	Interval	count
7350f74e8312d3d5add47d70f0bc8e91	Ð·Ð°ÑŀÐ±Ð¾Ñ‚Ð¾Ð°		1
6540a1730d818e544c93a7b0bdd61c9c	Ð·Ð°ÑŀÐ±Ð¾Ñ‚Ð¾Ð° 2022		1
899a8f292f10cd89d209726d65fcc476	Ð·Ð°ÑŀÐ±Ð¾Ñ‚Ð¾Ð° Ð½Ð° nft		1
52102bddb2df58ab594f226a0adcadaf	Ð·Ð°ÑŀÐ±Ð¾Ñ‚Ð¾Ð° Ð½Ð° Ñ‚ÐµÐ»ÐµÑ„Ð¾Ð¾...		1
e80ea779f8ba82bf5d5c10c51cd2c23e	Ð·Ð°ÑŀÐ±Ð¾Ñ‚Ð¾Ð° Ð² Ð¸Ð½ÑŠÐµÑŀÐ½Ð...		1
cb8df7784c4264263a81e0056689ccea	Ð·Ð°ÑŀÐ±Ð¾Ñ‚Ð¾Ð° Ð² Ð¸Ð½ÑŠÐµÑŀÐ½Ð...		1
d203f33bb6e9ec9d0238e29d4ccfc97e	Ð¸Ð³Ñ€Ñ‹		1
da2659d2571a04b01d297dec77ceb3be	Ð±Ð¸Ñ‚Ð¾Ð¸Ð½		1
61d1d5c36801d64dd82ecb0111b7da6b	Ð½Ð¾Ð²Ñ‹Ðµ nft Ð¸Ð³Ñ€Ñ‹		1
5577077b091b33069b9ac6e219ad9b59	Ð½Ñ„Ñ, Ð¸Ð³Ñ€Ñ‹		1
1b88d80ac91486fa096c12bd425f62a0	Ð½Ñ„Ñ, Ð¸Ð³Ñ€Ñ‹ Ð±ÐµÐ· Ð²Ð»Ð¾Ð¶ÐµÐ½Ð...		1
9bd4bd6934a7372a81de9aaaec5ebd65	Ð½Ñ„Ñ, Ñ‚Ð¾Ð°ÐµÐ½		1
8e9a9cdf46904a674a6ae6c826ea50e1	ÐºÐ°Ð° Ð·Ð°ÑŀÐ°Ð±Ð¾Ñ‚Ð°Ñ‚ÑŒ Ð½Ð° nft		1
9fc450226f2c5852a16e14a08786ad33	ÐºÐ°Ð° Ð·Ð°ÑŀÐ°Ð±Ð¾Ñ‚Ð°Ñ‚ÑŒ Ð² Ð¸Ð½Ñ‚Ð...		1
e8172e43371ffed953f188a3670d1214	Ð°ÑŀÐ¿Ñ‚Ð¾ Ð¸Ð³Ñ€Ñ‹		1
e39c177717ce8bfeeb0dae390ed772a9	Ð°ÑŀÐ¿Ñ‚Ð¾Ð²Ð°Ñ‚ÑŽÑ‰		1

Section 3: Network Visualization

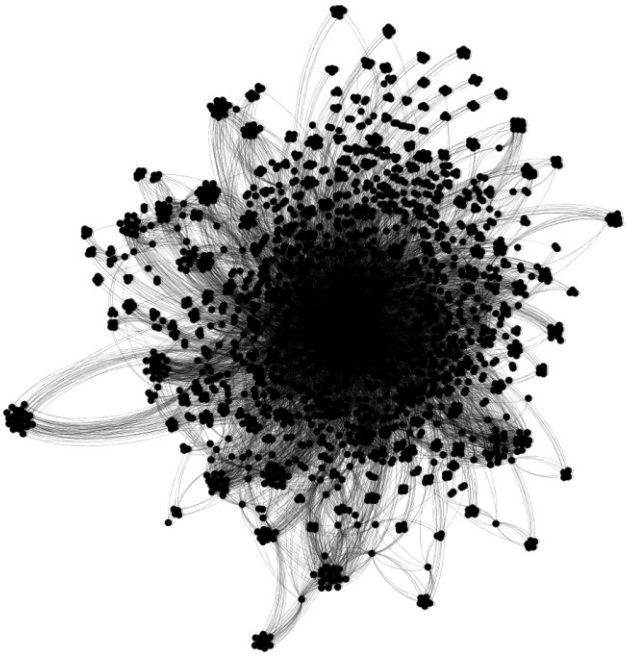
Visualization Process

For the Network Visualization we are going to present step by step the configuration in Gephi , and how the network looked in each step .

1) Initial Network



2) Force Atlas 2 – Default Settings

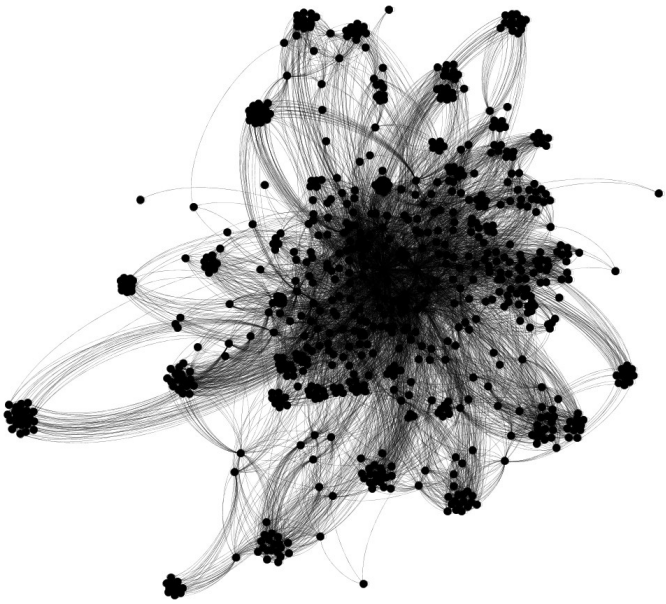


After running the initial algorithm Force Atlas 2 we can see a very dense center , which is to be expected after all the nodes (Tags) , are connected to the center tag crypto . We can also note some cluster's that start to form , and some tags that are weakly connected in terms of degree and clutter the view instead of providing information .

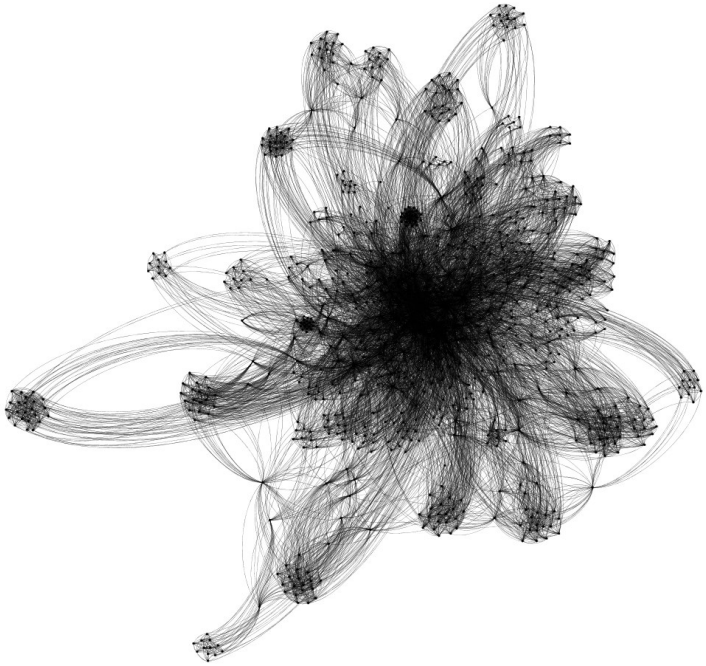
Step 3 : In order to clear out and filter some nodes out , we calculate the average degree and with the help of Gephi , we used a Degree Range from topology to do the necessary filtering .

Step 4: We also adjusted the layout a bit , increasing scaling to 50 from 2 and preventing overlap .

3) Filtered on Degree



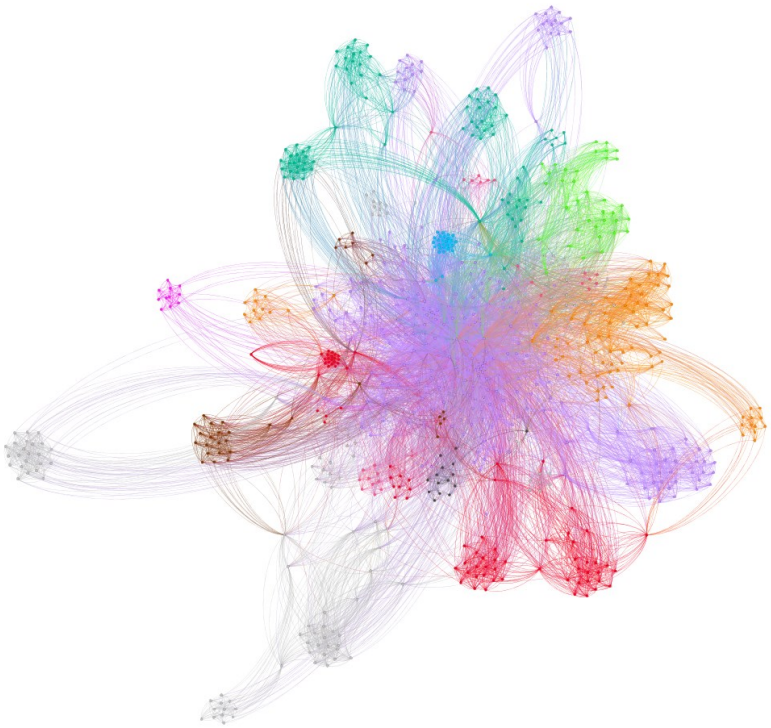
4) Adjusted Layout



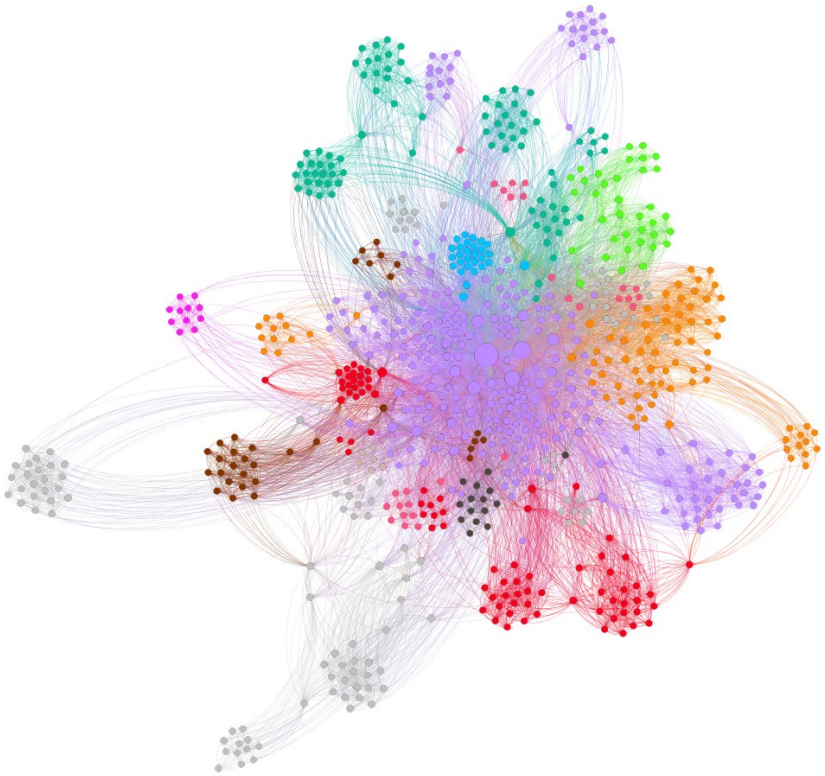
Step 5 : To better identify the communities of the graph , we use modularity to colorize it .

Step 6 : We also used Eigenvector Centrality to differentiate the size of the nodes , and increase the min size so we can get a better look at the individual nodes . The reason we used Eigenvector , is because we wanted to find the most important “outer” nodes ,(tags) that are also connected to really important nodes , like Ethereum , Nft’s , blockchain etc. , in comparison to using Degree for sizing we do not really care about well connected nodes , in their isolated communities .

5) Modularity Partioning

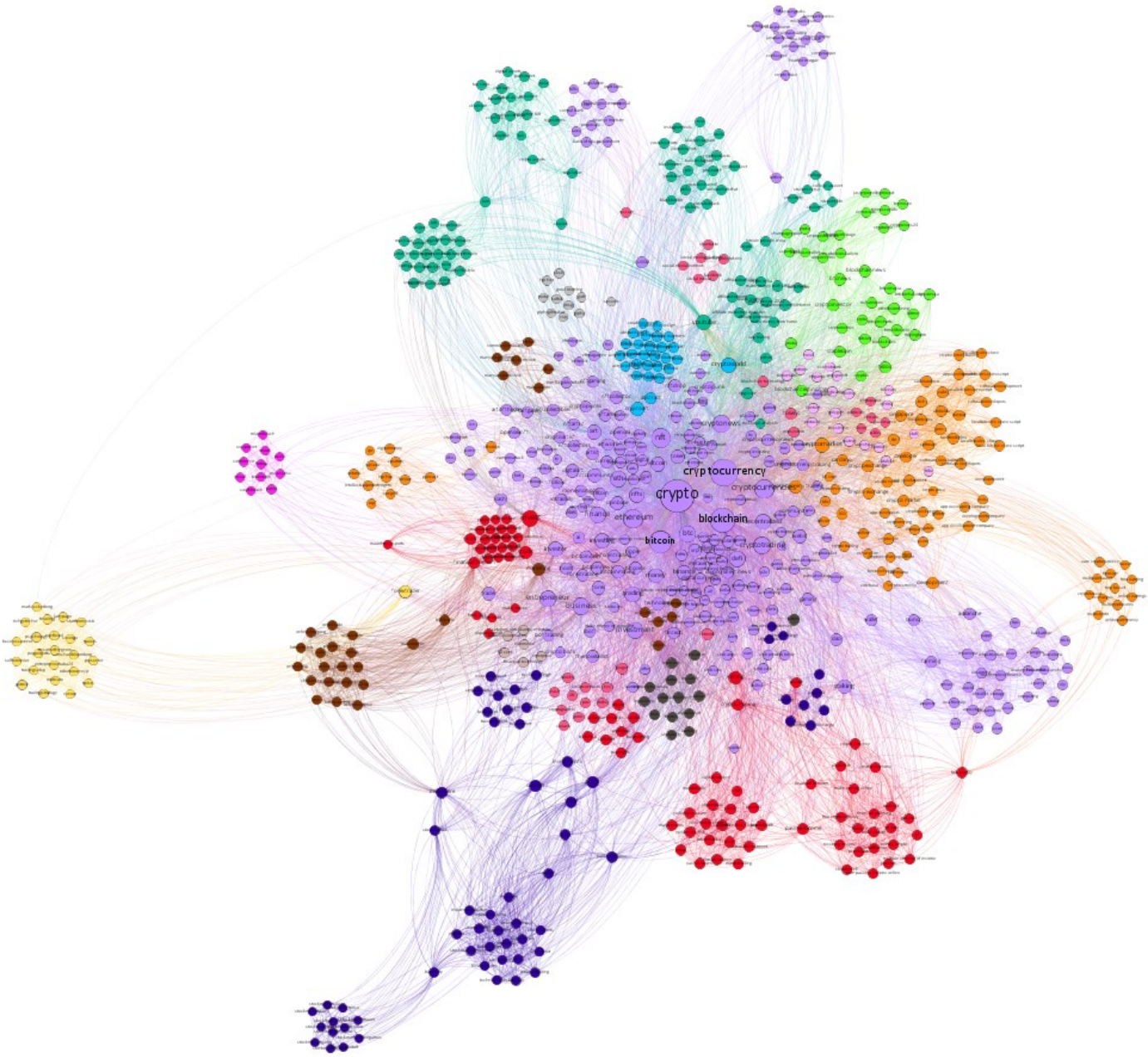


6) Sizing on Eigenvector



Complete Network Visual

After some final touchups , on labels and font size , edges opacity , changing some communities colors to avoid confusion , and nodes outline , this is the final network visual we get .



Section 4: Network Analysis

Topological Properties

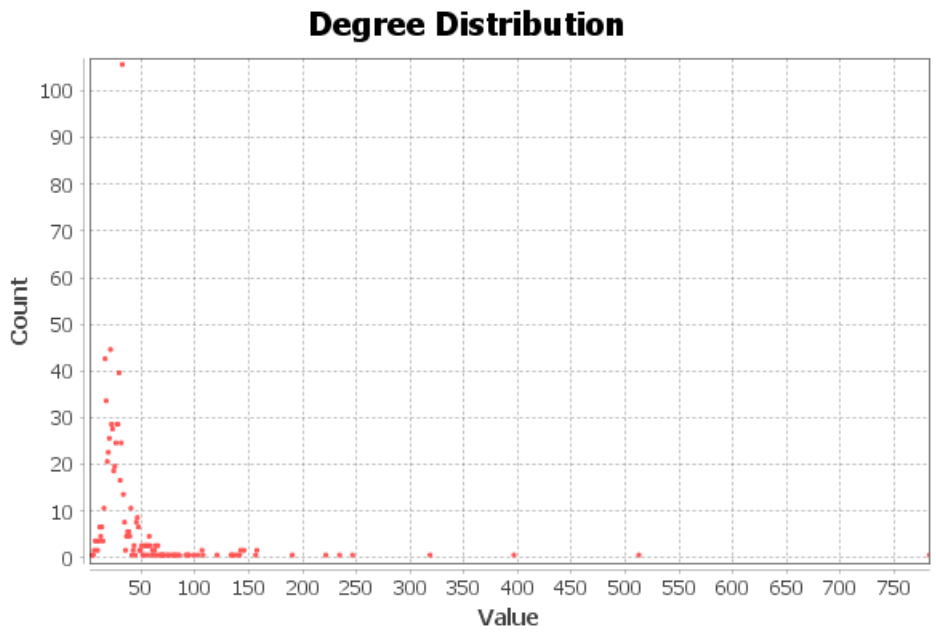
On our undirected information network, we have a total of **780 nodes (Tags**) connected by a total of **12925 edges** representing the co-existence of two tags in the same blog . The network diameter meaning the shortest distance between the two most distant nodes in the network is obviously going to be 2 , due to the connection of every node of the entire network to the central node ‘crypto’ , so even the most distant nodes are separated by 2 edges . **The average path length** , meaning if we took all the possible combinations for pairs of 2 out of the 780 , (780 choose 2) for a total of 303810 combinations , summed their shortest paths and divided by the total amount of possible combinations , the result would be **1.96** . The closeness to 2 can be explained as most node combinations have a path of 2 because like I mentioned before we have that huge central node ‘crypto’, but we also have to take in account the 780-1 = 779 , direct connections from every node to the central node , that have a path of 1 .

Component Measures

Again the central node that all the tags are connected to makes it obvious that there is **only one huge component consisting of the entire connected graph**.

Degree Measures

Degree is the number of edges incident on node , more simply put for an undirected network the “lines” , we can see attached to a node . For our network the **average degree is 33.141** , which means that on average a node is connected with 33 other nodes . The highest degree can once again be found on the central “crypto” , node with a degree of 799 , other nodes worth mentioning with really high degrees are cryptocurrency , blockchain , crypto news , nft .

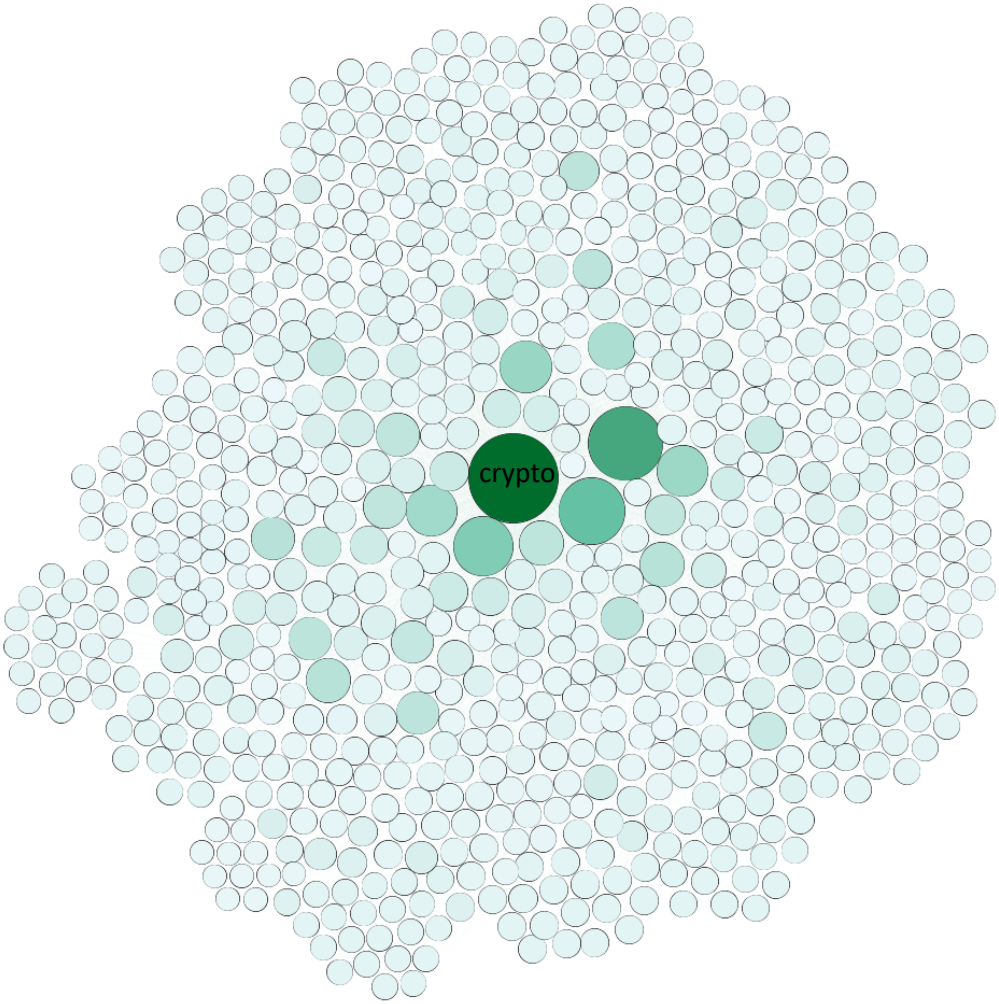


The minimum degree which we created “artificially” by imposing a degree range filter is 15 and the highest one can be found on the central node “crypto” with a degree value of 799 .

Due to the extreme range of the distribution , a visualization on degree ,wouldn’t have some value , below I used stronger gravity so the more distant nodes wont get lost in the white space , to demonstrate how the extremely high values of the central nodes and its close neighbors , render the rest the majority of nodes insignificant .

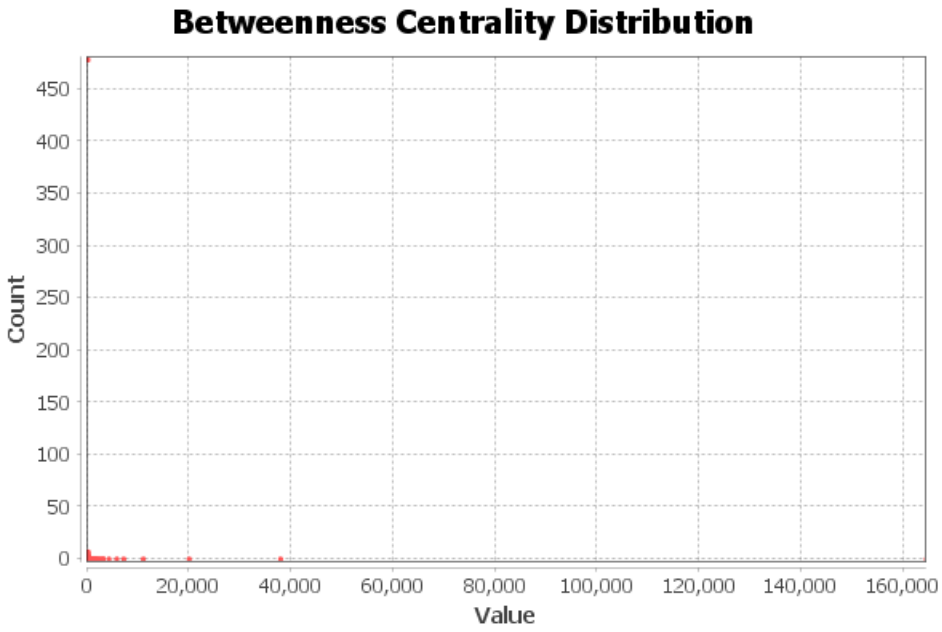
Note , White to Green from lowest to higher degree values , and the sizing remains ranked on Eigenvector. Primary Nodes we can note that are deeper greens are , crypto , btc , cryptocurrency ,cryptonews ,Ethereum , nft .

Terms we tend to hear most often , in the community .



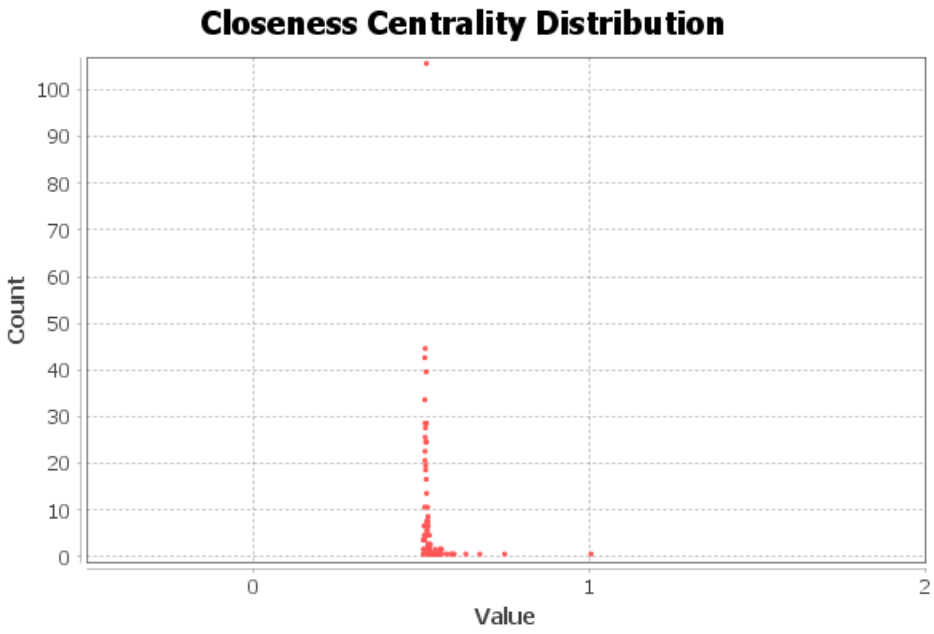
Centrality Measures

Centrality measures answers the question , about how central is a node to the network . Let's start with **betweenness centrality** , that depicts how important is a node in terms of connecting other nodes through the initial node , more simply put the number of times a node lies on the shortest path between other nodes.



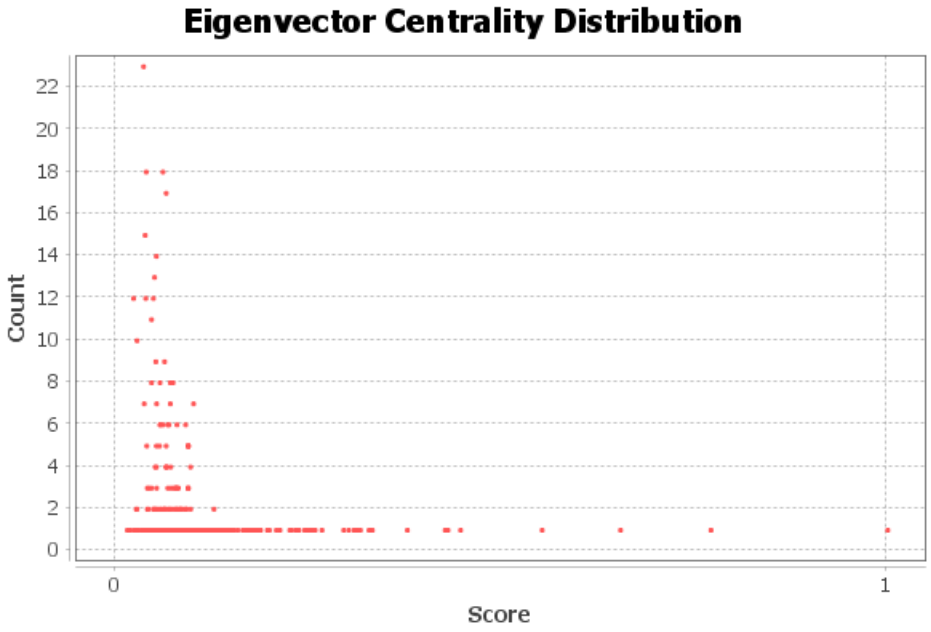
If we zoom in we can see that the majority of the nodes gather around really small values for betweenness centrality , with the most important node being once again the "crypto" . with a **range from 0 to 164382**. A utilization of between centrality would be to find link's between different communities of tags , in other words what nodes act as step between communities of tags .

Continuing to **Closeness Centrality** that indicates which nodes best placed to influence the entire network most quickly , by calculating the shortest paths between all nodes, and then assigning to each node a score based on its sum of shortest paths. In our network the closeness centrality ranges from 0.500321 to 1 with the following distribution .



It’s pretty obvious that the node we would choose to effect the whole network is the central one , but what is interesting is that all the other nodes have kind of similar value , what could possible be valuable is to isolate each community and find the most important node for each cluster by isolating said cluster , and magnifying the small differences.

Next we are going to measure centrality using Eigenvector , which we have already calculated to make the more important according to it nodes bigger in our visualization , Eigen centrality can identify nodes with influence over the whole network, not just those directly connected to it , by giving more weight to nodes that have connections with nodes that are better connected themselves .

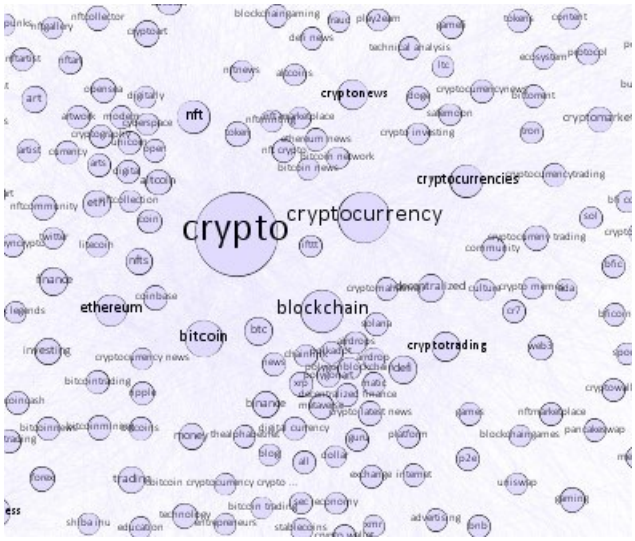


As we can see the values range from slightly above zero , to the maximum of one for the central node , with the majority of the nodes to be found in the 0.1 to 0.3 value range . Instead of the usual nodes , eigen vector recognized as important nodes-tag like defi (decentralized finance) , decentralized , games , crypto art , and Ethereum that the closeness and betweenness metrics failed to do so .

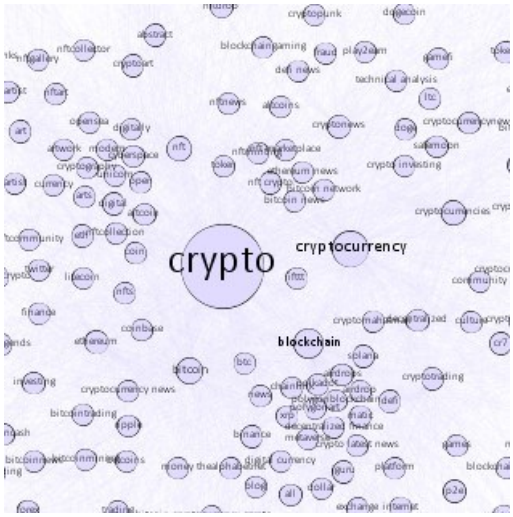
Different Visualizations on Centrality Measures

Figure 1 has sized by ranking on Closeness Centrality , Figure 2 has sized ranking on Betweenness Centrality , both have been zoomed to the central cluster of the network , because it is the only place of the graph where the ranking creates differences based on the metric with which it has been sized .

(Fig 1)



(Fig 2)



I choose Eigenvector because it was the one that offered the most insight , and differentiation about centrality ,from the Betweenness and Closeness Measures we got some pretty obvious results , about which nodes are important.

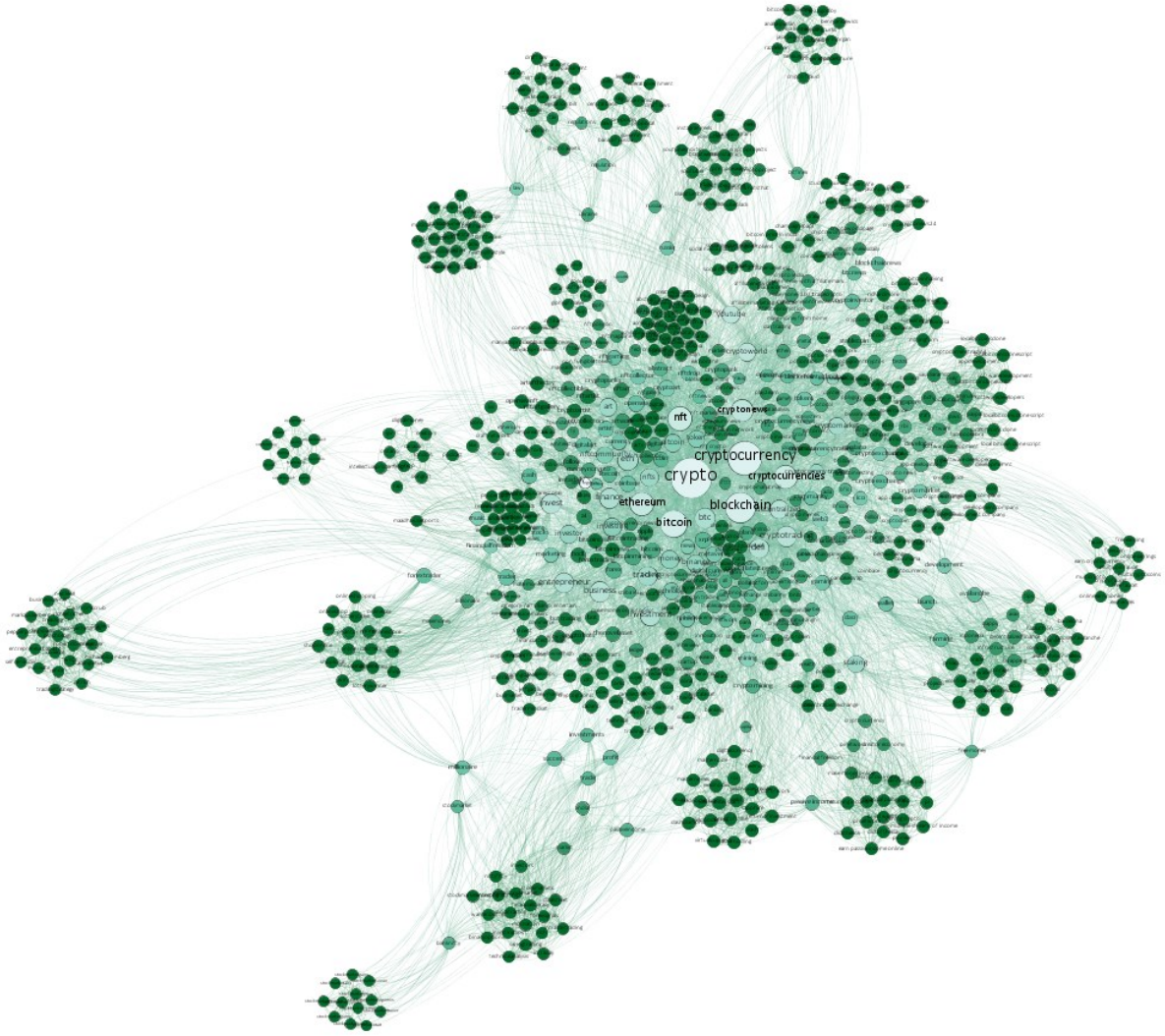
Clustering Effects on the Network

Let's start with the average clustering coefficient , that tries to capture to what degree the of our network tend to cluster together . The value range from zero to one depicting how tightly connected they are , zero for a star , one for a tightly connected "clique" . In our case the **Avg Clustering coefficient is 0.85** indicating a tightly connected clustered network .

From the distribution we can understand and, in a way, predict what the visualization has already showed us , there is a pretty high clustering to be found in the network in fact we can also see perfect **scores of 1 , meaning there are cliques forming in our network** , while also the central cluster that surrounds the principal node is spread out , and finally some "singled out" nodes falling between the white space that act kind of like connections between the existing clusters , with really low scores .

Moving on, **total number of triangles is 102897** , it's a pretty good measure of the cohesiveness of a network , what it actually counts is cases where a set of three nodes where each node has a relationship to the other two forming quote to quote a "triangle" , in graph theory it is sometimes referred to as a 3-clique.

What we could do to identify where triangles form , where the network is most tightly clustered , and perhaps locate cliques is to better visualize this , we will create a version of the graph ranked on clustering coefficient .

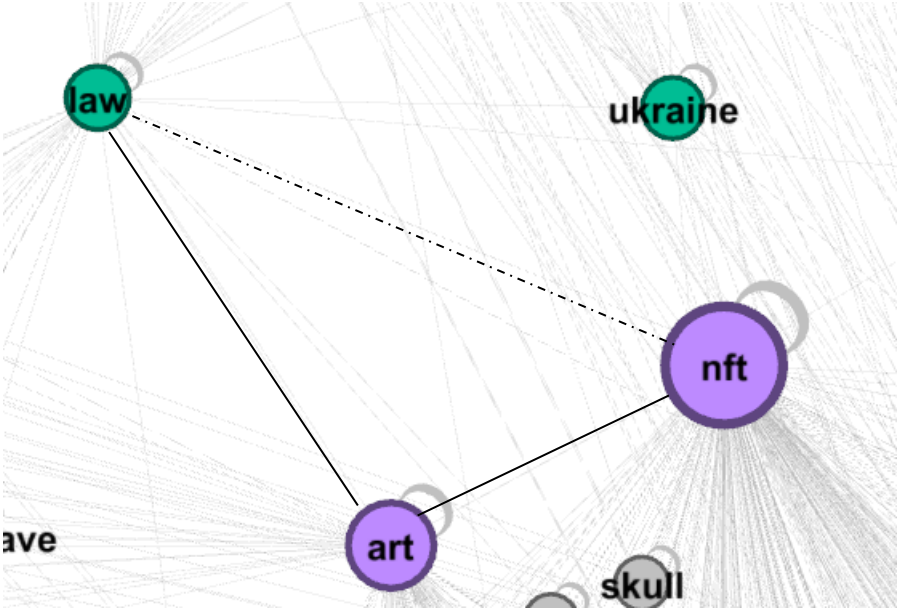


What we can note from the above , is deep green parts of the graph for high connected subparts of the graph where we would find the majority of the formed triangles, and lighter green for nodes that act as link between those sub-parts .

Next , we are going to examine cases of the **triadic closure phenomenon in our network**. What the triadic closure phenomenon essentially studies is how a network evolves dynamically through time , how the basic components of it , nodes and edges form and vanish as time progresses , what it suggests is that if two nodes are not connected directly but they do share a connection with a mutual node , then the probability that they are going to be directly connected sometime in the future is much higher .

Below we are going to zoom in and find some cases from our graph .

We found something really interesting , to exhibit a **triadic closure** . Art is connected to the tag law , a connection that is also present in the real world , art is regulated and has legal code , on the other hand there is a pretty deep connection between nft's and art , in reality digital art is the only application of nfts thus far in the real world . This implies a future possible connection , that is bound to happen , regulation and a legal code is very much going to be brought upon nfts after the huge amounts of transactions we see happening in the last year .



If we wanted to utilize the **strong triadic closure property** , we would weight the connections based on the amount of times they co-existed in a blog . An example of strong triadic violation we found in our network is this case . The tag binance referring to a well known trading platform is connected pretty heavily to both bitcoin, and to the tag crypto exchange but to our surprise there is no connection bitcoin-cryptoexchange.

Bridges and Local Bridges

Given the nature of our network **, we do not have a single bridge** in our network , because the term of bridges means that if we were to disconnect two nodes , they would lie in a different components . In our network even if we were to disconnect a node with the central node , it surely has connections with other nodes that in turn have a connection with the central node , that connects with the entire network , making the existence of a bridge impossible.

Something that is much more common in real life social networks ,are local bridges . The definition of a local bridge , for example for A and B two nodes that are connected is if the endpoints of the two nodes have no friends in common , and the dismissal of their original connection would increase the shortest path from A to B .The distance of that new shortest path after the initial edge has been deleted , is called local bridge span , and the initial connection a local bridge.

Once again due to how our network is connected no matter what connection I to try to remove , the two nodes will always have a friend in common the ,central node “crypto” , so it will always break the rule on being a side to a triangle . For demonstration purposes I am going to remove temporarily the central node.

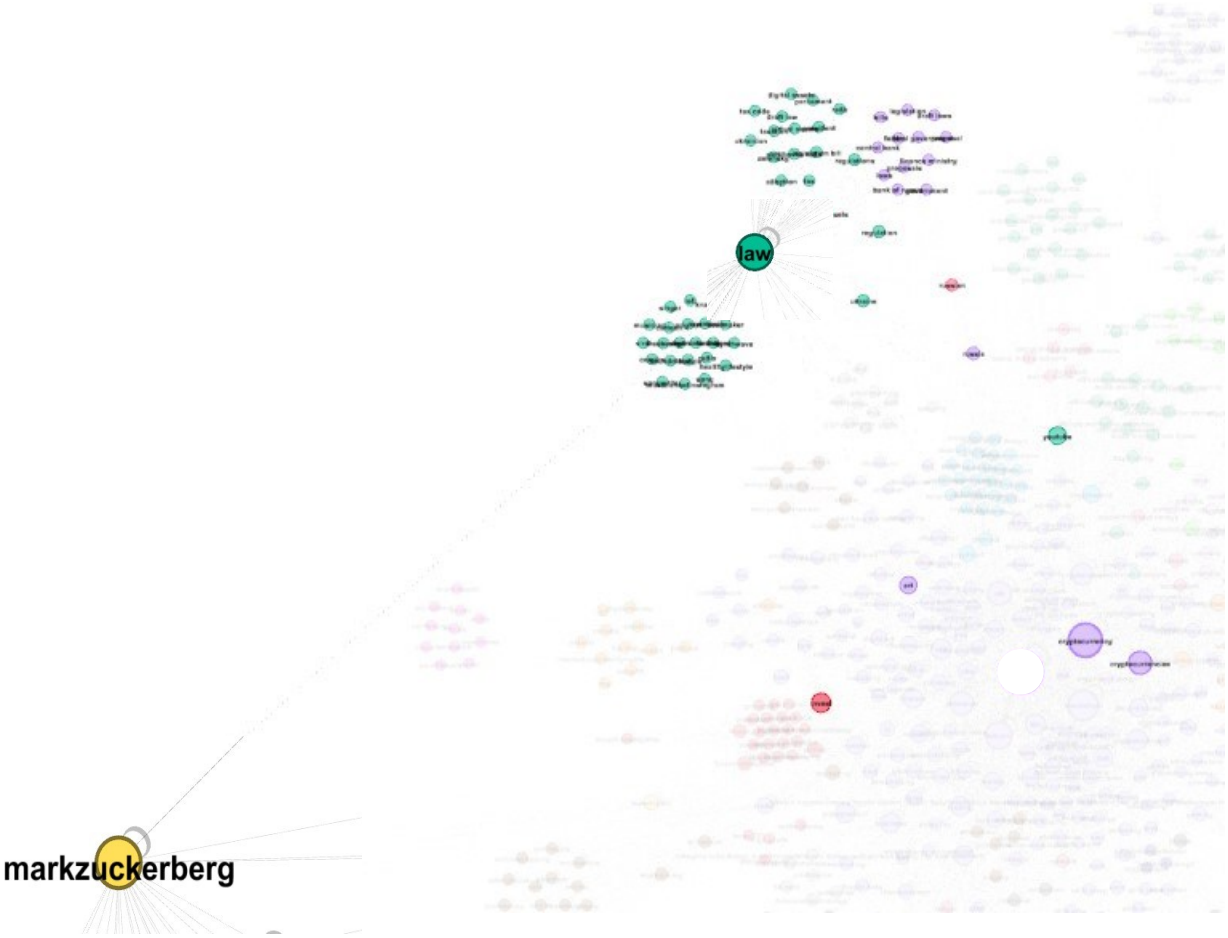
I also used the bridging Centrality measure , to identify nodes that could potentially have edges that would be bridges , but even with the crypto node removed the values where extremely low and nothing of value came up .

Local Bridge Example

Here we have an example of two nodes law and lizard , being connected , the image above is from the lizard's connected nodes point of view , after this we will also showcase a snapshot of law's connected nodes to demonstrate that the two nodes do not have common neighbors .

Tag law connected nodes.

The graph is zoomed to the two points of interest



If we were to disconnect that edge , the **new shortest path so the span of the local bridge** would be law to invest , invest to forextrade , forextrade to markzuckerberg so a span of 3 for a new shortest path .

Homophily

The homophily theory states that nodes with similar attributes and personal characteristics can strongly predict the existence of the connection between them . What this suggests is that there are factors beside the basic principles , edges and nodes, of a network that effect how the network is structured and how it will continue to evolve. What this also could suggest is perhaps , the ability to predict is whether two nodes are connected based on the level of similarity in features I choose to compare . For example in a network of humans , race and gender could be used for a homophily analysis . In our case we do not have such features for each node , what could be of interest is if we had available categories to each tags like technology , art , economics, and using that to look into Homophily . Instead **what we can do** is check for Homophily based on the amount of occurrences generally of a tag in blogs , so we could see if there is an increased probability of a connection between words that are tagged the most , and generated metrics like degree and centrality measures..

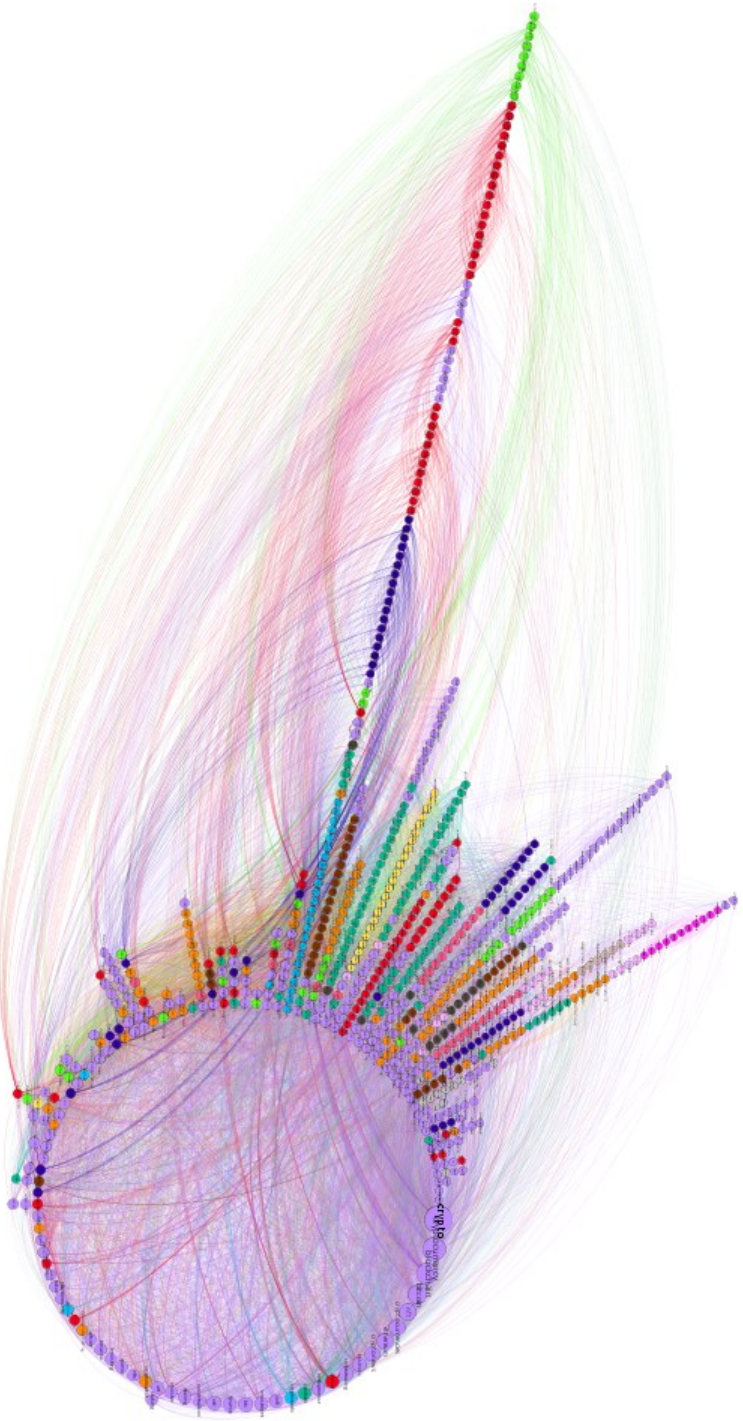
To study homophily in Gephi , we are going to use the Radial Axis Layout provided by the Circular Layout Plugin , by finding distributions of nodes inside groups with their links . We also enabled the Draw as Spiral to make it more readable

Grouping by node count , failed the layout was essentially a straight line .

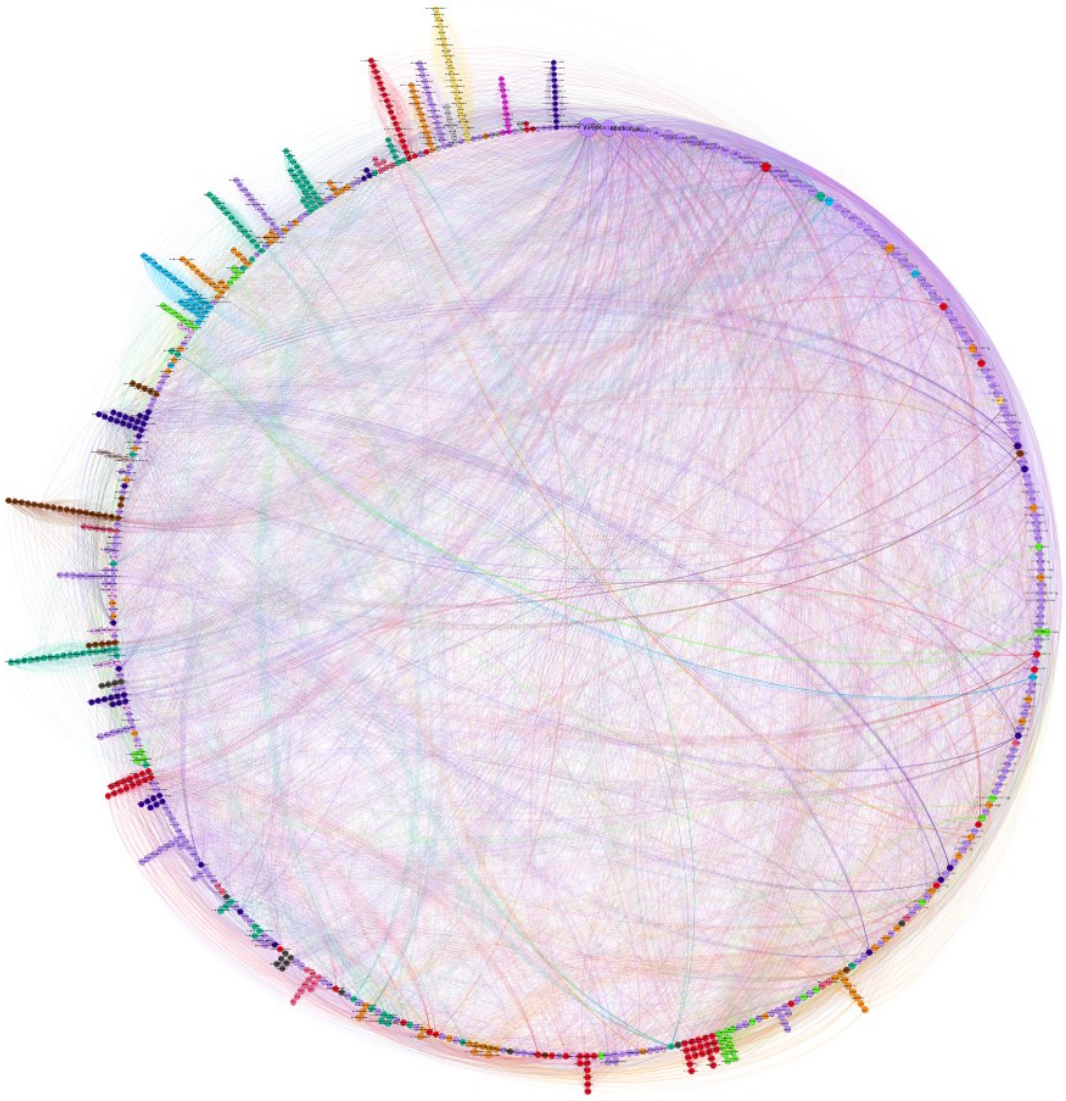
Grouping by Eigenvector , showed no essential grouping , we had an almost perfect circle , but the few spots that had a level of grouping , showed very high connectivity actually it managed to identify groups that were cliques , supporting the theorem of higher change of connections between groups .

Grouping by Degree , kept to some level the communities we have created intact (more details on communities on the next chapter).

Group by Degree



Group by Eigenvector



Small World

Our network definitely fits the “small world” model , an idea that contemplates that things , person , information from real life social networks is much closer and connected than we think . For our network the avg path length , meaning the shortest path connecting two nodes – tags in is only 2 steps away . Some kind of peculiar connections we can note in our network , hip-hop to the president of Ukraine , Michael Bloomberg the politic to a well known nft project called cryptopunks , comics to larry page the founder of google , only 2 steps away . If we were to remove the central node once again , how much would the avg path length increase to , would it make us deem our network unfit to be called small world ? Let’s rerun some metrics without the central node , the diameter increased to 4 and the avg path length to 2.2 from 1.98, so all the unexpected connections we mentioned before even without the central node that made things easier , are only 4 steps away maximum.

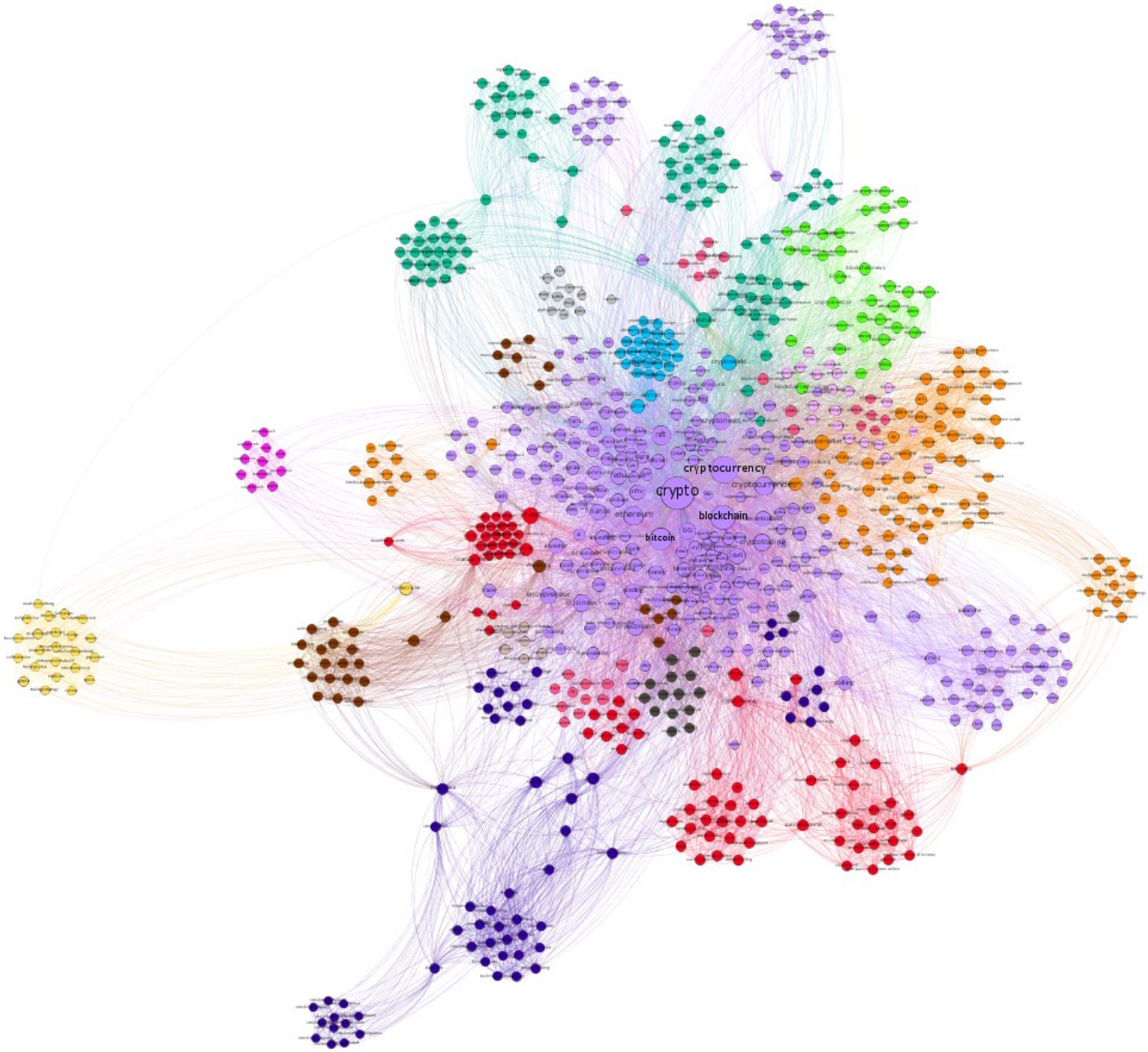
Graph Density

The term network density represents the proportion of actual “relationships” meaning the existence of an edge between nodes , in comparison to the total number of possible edges . For example if we have an undirected network with nodes (A , B , C) , all nodes can be connected with a total of 3 edges AB , BC ,CA , but in reality we could have a linear AB , BC that ratio of actual to possible connections from 0 to 1 indicates how sparse or dense a network is . In our network d for **density is 0.043** like almost all real networks , its sparse but not to an extreme , but in comparison to other prominent networks we could say it’s pretty dense , but that could also be due to the initial filtering we did on degree.

Modularity and Cliques

Modularity measures the structure of a network , its ability to divide itself into more compact and dense communities , a network with high modularity would have really dense parts of highly connected nodes , and weaker connections between nodes of different divided communities .

In our case with the help of Gephi (Randomizer on for better results and utilizing edges weight) we calculated **Modularity for a value of 0.427 with a total of 15 communities** . I am reposting the visualization with the color division based on modularity to make it easier to follow along with the description of each different community .

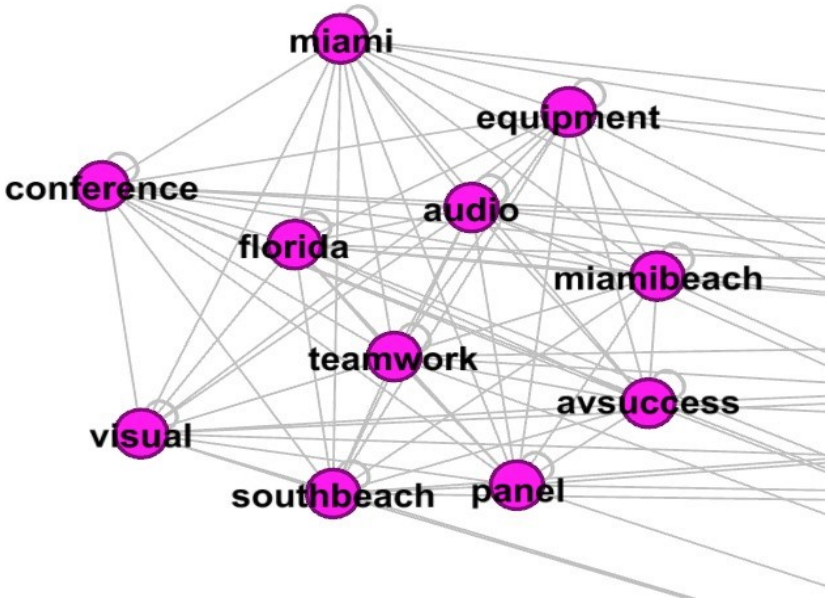


- Community 1 Light Purple : The biggest community , covering 32.95% of all nodes , containing the central node-tag crypto , and has all the most known and used search terms and tags around cryptocurrencies like , blockchain ,bitcoin , decentralized , nfts ,Ethereum, altcoin .
- Community 2 Orange : Tied with the green community for second largest , covering 10.77% of all nodes , with tags around blockchain development , mining bitcoin , and earning crypto , the more tech savvy side of things .
- Community 3 Deep Green : Contains everything associated with the law , and regulation of the cryptocurrency world , tax bills , tax code , and the instruments that apply them . It also through regulation , contains something unprecedent a tags related with making money from online socials , and tags about Ukraine and President Zelensky through parliament .
- Community 4 Deep Red : 10.51% Internet Entrepreneurial motivation scams , on trading and making money through crypto for financial freedom .
- Community 5 Deep Blue: 7.82% , Warren Buffet , Day trading, technical analysis , everything associated with the stock market.
- Community 6 Neon Green : 4.82%: Tags containing the word news , on technologies related to blockchain , financials of well known coins , drake , and the democrats ,Weird connections .
- Community 7 Brown : 4.23% Online sales , e-commerce , sales.
- Community 8 Teal : 3.59% Creative art , digital art , motion design , and creative code . Tags around digital art that is becoming more prominent the last year through nfts.
- Community 9 Light Yellow : 3.05% Celebrities , and motivational quotes .
- Community 10 Deep Purple : 2% Miami , and South Beach probably because conferences are held there for crypto enthusiasts.

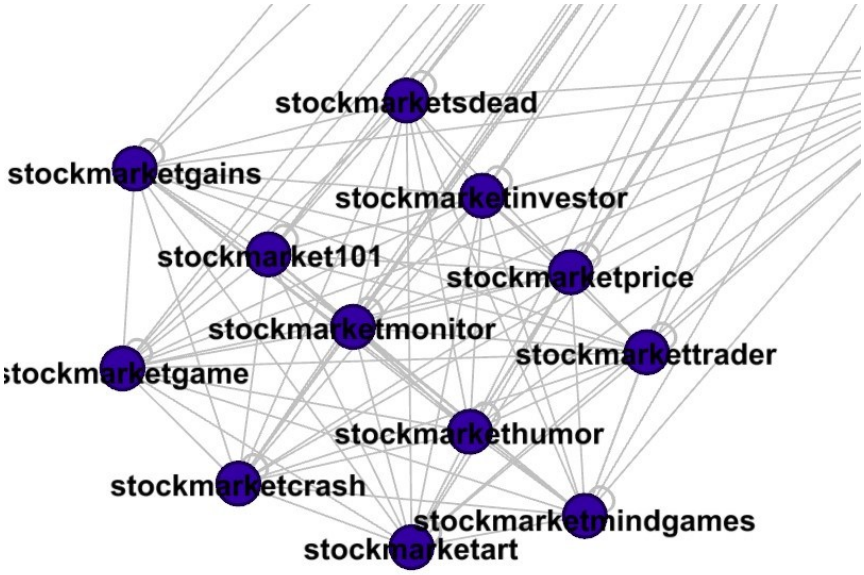
  Other communities didn’t provide useful information , I believe that because we took the last 2000 blogs , and not blogs say for the last 6 months it was bound to have some noise , and random connectiveness that doesn’t amount to something valuable.

Some Examples of Cliques

The definition of clique is a fully connected subgraph , fully connected means that all possible connections are existent , in a small part of our whole graph , everyone is connected to everyone .



A clique that also happens to be a community by itself , about the east side and crypto conferences.

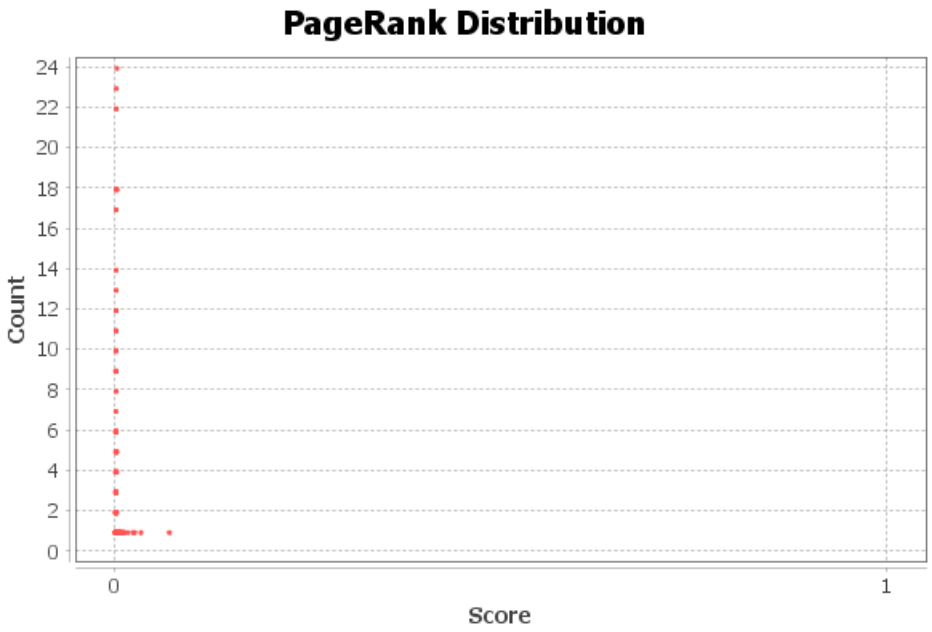


A clique about the stock market.

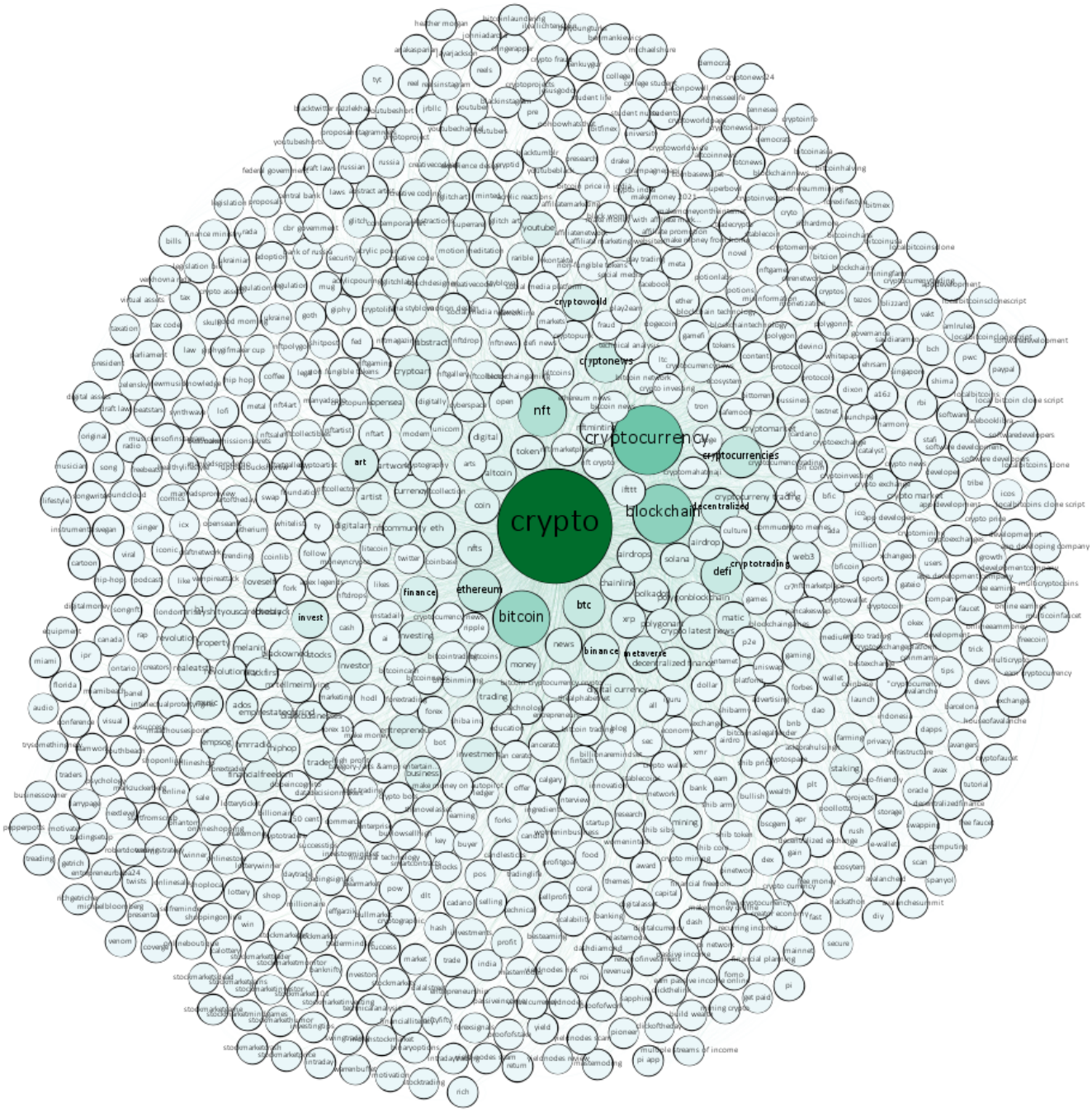
PageRank

PageRank is a variant algorithm of the Eigenvector , that is used like all the other notions of centrality , to capture the importance of a node in terms of centrality to the whole network . It was used to capture the importance of pages by Google, considering the amount of links that pointed to a page and also the quality of the pages that pointed to the page ,(the PageRank of the pointers themselves) , from the words pointer we can understand that **PageRank is primarily used for the centrality of directed networks** . Unfortunately we have an undirected network , but we run the algorithm nonetheless to see how it pans out .

For our network , we run PageRank with the default settings on probability and stopping criterion , but with a slight difference of using a edge weight as a criterion .



Something peculiar we can note is the vast differences in values and distribution from the Eigenvector centrality on which the PageRank is based on , below we are going to present a visualization where both the nodes size and coloring are based on their PageRank values .



Closing with PageRank , we can notice that is returns as most important , the usual nodes of cryptocurrency , bitcoin , blockchain , and nft . Like we mentioned before the algorithm is not optimal for undirected networks which is why we didn't use it in our initial visualization.