Analyzing Food Network and Implications

Deepkumar Patel
DATA-51000-001, Summer 2020
Data Mining and Analytics
Lewis University
deepkumarpatel@lewis.edu

I. Introduction

The raw dataset included in this report contains mutually liked Facebook pages from the food category, it was collected in November of 2017. It contains two different files, specifically, the nodes and edges. In this dataset, nodes represent the pages, and edges represent the mutual likes by users. In addition, the nodes also include labels, which represent who the page belongs to, essentially a page name. The dataset is publicly available on the network repository website under the social networks category [1].

The purpose of this report is to analyze the food pages network and detect the community within them. Using this information, we can find common interests between the users. Network analysis has many implications, for instance, the knowledge that is gathered can be used as a recommendation system. We can even take a step further and use this information for the purpose of advertisements on different platforms. In this report, we will use two different methods to visualize our network, namely, multilevel agglomerative edge bundling method [2], also known as Yifan Hu in Gephi, and graph drawing by force-directed placement method [3], also known as Fruchterman Reingold in Gephi. The analysis and different methods which we will see in future sections are no way limited to this dataset, therefore, we will use the food network dataset as an example and generalize the idea for any type of network.

The future sections of this report describe the data preprocessing, dataset, two different methodologies, results gathered from both methods along with a discussion, and a conclusion. Annexed is a list of sections with a more detailed description. Section II describes the process that is involved before any type of network analysis is done on the dataset, this includes correcting and formatting the dataset in a format Gephi can understand. Section III describes the dataset. Section IV describes how the multilevel agglomerative edge bundling helps alleviate cluttering by revealing high-level edge patterns. Section V describes graph drawing using force-directed placement to reflect symmetry. Section VI compares both the algorithm mentioned above, along with the discussion of the results. Section VII is the conclusion of this report.

II. DATA PREPROCESSING

The original dataset contains 620 nodes or vertices and 2,102 edges or links. It is an undirected and unweighted graph with few self-loops. It is also worth mentioning that the pages or vertices included in this dataset are blue verified by Facebook. It means the pages are authentic or the pages are of public figures, media companies, or brands. They are often searched pages on Facebook. For the purpose of this report, we will use Gephi as our main tool for analyzing the network and Jupyter notebook to create plots, both of these files are included with the submission of this report. To further understand how this dataset was prepared and transformed, let's take a look at the following steps. In addition, also keep in mind that the following steps are only needed for Gephi, other tools may have different approaches.

A. Update Headers

In the fb-pages-food.nodes file, "new_id" is renamed to "Id" and "name" is renamed to "Label". In the fb-pages-food.edges file, I added headers to both the columns, the first column is called "Source" and the second column is called "Target".

B. Add Type Column

For both the files I added a new column called "Type" and its value is "Undirected". The reason we have to do this step is because Gephi would consider the network as directed graph. Since our network is undirected, we have to explicitly specify values as undirected.

C. Drop Unused Column

In the fb-pages-food.nodes file, drop the first "Id" column. The reason why this column is there in the first place is not made clear by the authors. It also doesn't provide any value to our analysis; therefore, it is not needed.

D. Identify and Correct Label Values

The label column in fb-pages-food.nodes file contains values such as "???", "???? Tsui Wah Restaurant", etc. Therefore, remove any question marks or special characters which do not make sense to be a part of a label. It also contained blank or null values, which gets replaced with string "Unknown". The reason for not deleting nodes with null labels is because it might be connected, or be part of a bigger community.

E. Export Nodes and Edges

This step is self-explanatory, in this step we just export both the files as comma-separated CSV.

III. Data Description

All the steps mentioned in the previous section will take care of smoothly visualizing our network in Gephi, however, in this section, we will see how even all the pre-processing is not enough, as we will be applying more filters to make our network more manageable. After the data preprocessing step, we are left with the following attributes shown in Table I. Since we have two different files, a new file column is added to this table, which simply shows the associations between a file and it's attributes.

TABLE I: Nodes and Edges Attributes

Attribute	***Type***	***Example Value***	***Description***	***File***
Id	Numeric (primary key)	386	Node or page unique identifier	Nodes
Label	Nominal (string)	Josh Marks	Entity the page belongs to	Nodes
Type	Nominal (string)	Undirected	Graph direction	Nodes
Source	Numeric (foreign key)	0	Source node or page unique identifier or a foreign key of a source page	Edges
Target	Numeric (foreign key)	276	Target node or page unique identifier or a foreign key of a target page	Edges
Type	Nominal (string)	Undirected	Graph direction	Edges

When we initially load the nodes and edges file into Gephi, we see a very complex and un-manageable graph as shown in Fig. 1, it has 620 nodes and 2,102 edges. In order to make some sense out of this network, we will use pre-installed filters in Gephi. As mentioned earlier, this network is just an example, therefore, the following steps can be applied to any type of network, which means you would have to tweak the settings for your own purpose.

A. Giant Component

This filter is located under the topology category. By applying this filter, any node that is not connected with the main cluster will get removed. The reason for this step is to get rid of any disconnected nodes, as these nodes don't tend to contribute too much to the main analysis. Having said that, it is also a good idea to look over those nodes, as possibly they do somehow affect the other cluster of nodes. In this network, we don't see any disconnected nodes or graphs, therefore, we don't have to apply this filter.

B. Change Layout

In order to see what the network looks like, I find that Yifan Hu algorithm works best to visualize the network. We will study more about this algorithm in the later section, but for a quick observation, we use this algorithm. After running this algorithm, the result we get is shown in Fig. 2, which clearly shows the clusters. However, if we carefully observe, we can also see there are nodes (around the network) that are connected to many nodes with a degree of 1. Essentially, a hub with many nodes that has a degree of 1. One of the reasons we might see such a pattern is because the data was never collected beyond those points. In terms of our dataset, those nodes would still provide us some value as our network is undirected. However, that is not the case with every real-world example. For instance, assume we have a network of emails, where each node represents an email and edges represent to whom an email was sent to. In this case, once we reach an ending node, there's no other way for us to traverse back, as email networks are directed graphs. Therefore, in cases like these, we could remove those nodes to reduce graph cluttering. In order to achieve this goal, a possible solution is provided in the next step.

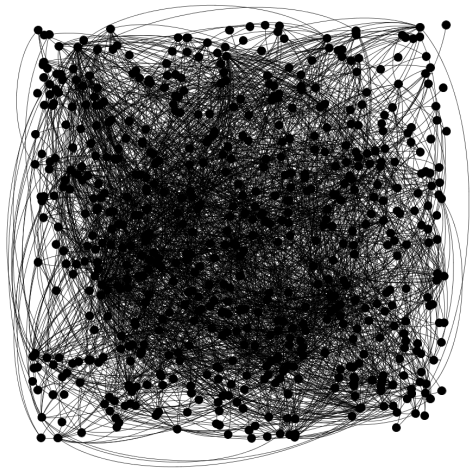

Fig. 1: Initial network of food pages (representing as nodes) and its edges

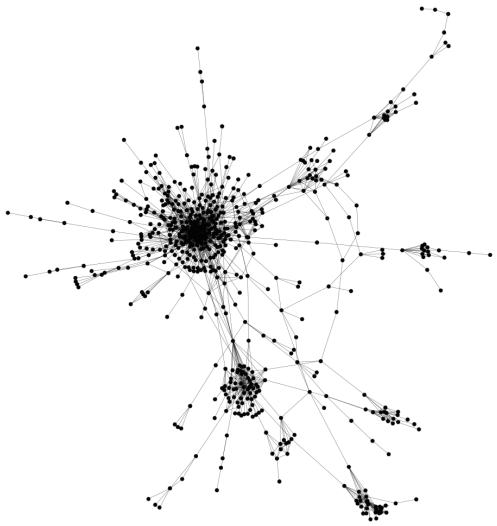

Fig. 2: Yifan Hu algorithm layout

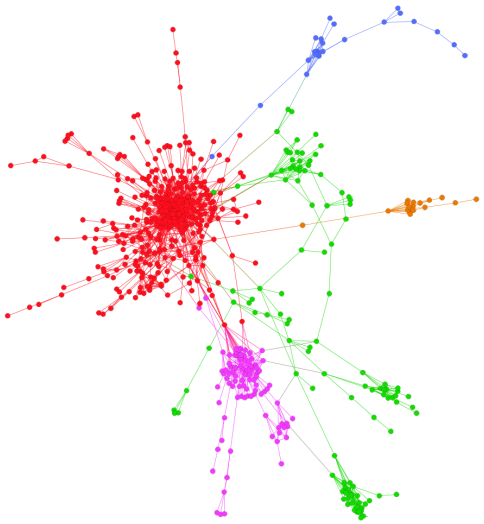

Fig. 3: Communities within food network

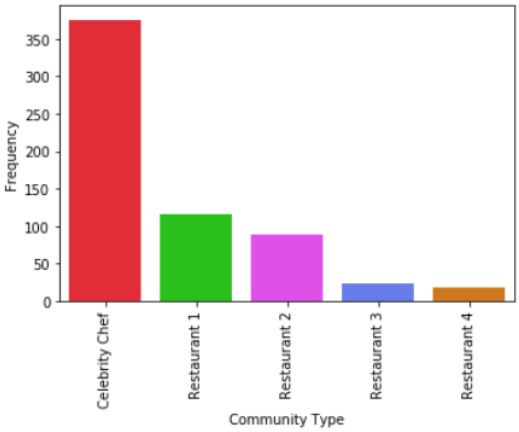

Fig. 4: Community type and frequency

C. Degree Range

This is just an extra step which can be used to remove nodes with a degree of N. This step is useful for directed and undirected graphs. In the later sections, we will see how this filter will come in handy.

D. Community visualization

In order to better understand the food network, we will apply a community detection algorithm called modularity, it is provided as a simple statistic setting in Gephi. There are many networks that naturally divide themselves into communities. Using Fig.2, we can already see which communities are going to form. Picking a correct number of resolutions is also a challenge, but fortunately, we have a rough idea of the communities, therefore after multiple trial and error, the resolution of 8 was picked. In Fig.3 we can clearly see those communities, and for food networks, it created 5 communities. Now that we have the communities, we can categorize them by understanding it's label. In order to avoid label cluttering on the network itself, a new plot was generated using python. Essentially, all the nodes were exported along with its modularity values, then using python; group the data by modularity. After careful observation of all the labels, I came to the conclusion that all the small communities are the pages of different restaurants, and the main community in red, are the pages of celebrity chefs as shown in Fig. 4. To see all the labels per modularity class, please refer to the Jupyter notebook provided with this submission.

Now that we are able to see the distinct communities, we can further evaluate this network. There are two choices we have in this network. The first choice is to further evaluate the community pages of chefs, and the second choice is to evaluate the community of different restaurants. There are no advantages or disadvantages over either of the choices as both types of community will provide us with some useful knowledge. The idea is to find how pages are related to each other, and if they are related in some ways, why are they related? These are the type of questions one can ask for network analysis. For the purpose of this report, we will first examine the celebrity chef community by applying an edge bundling algorithm with different statistic measures. After that, we will examine the community of different restaurants by using force-directed placement to reflect symmetry, along with different statistic measures.

IV. MULTILEVEL AGGLOMERATIVE EDGE BUNDLING

One of the main purposes of creating this algorithm is to visualize large networks in a simpler form, and at the same time, reveal high-level edge patterns [2]. Data collected from the real-world are often compact and have encapsulated relationships between the objects, and it gets difficult to see any underlying patterns. In our example of the food network, we had a similar issue (Fig.1) where everything is jumbled into one giant network with no patterns. Even though our network is small in terms of size, it still suffered from the visual-clutter problem in visualization. There have been many algorithms, techniques, and filtering methods to reduce cluttering. However, edge bundling seems to be the most common technique. Essentially, this technique tries to group similar edges into bundles, in turn, providing an uncluttered view of a network. According to the authors of this research paper, this algorithm is much faster than the previous edge bundling technique, as it uses a multilevel agglomerative edge bundling method. To better understand this method, the authors gave a very simple intuition. Imagine what a human operator would do if there is a task where they have to bundle a large number of wires in terms of its length? First, identify wires with similar lengths, and if they are similar then merge them, otherwise, start its own bundle, after that we keep on repeating this process until we are done. Clearly, this is just an insight into how this algorithm works on a very high level, however it is much more involved. For more detailed and involved reasoning, the explanation is available here [2].

Before, we apply this algorithm using Gephi, we first have to get rid of all the restaurant communities, because this section will only cover the celebrity chef community. In order to select a specific community, we will use the modularity class filter, it is available under the range category. For our case, we only selected modularity class with value 0 because this is the celebrity chef community. After we apply this filter, we are left with Fig.5. Even in this small community, we have 376 nodes and 1,434 edges. At this point, we could observe all the labels and try to make sense out of them, but for this report, we will further apply the degree range filter. This filtration will not only remove nodes with low degrees, but it will also bring more clarity to our analysis. In an ideal situation, we should not get rid of any nodes, but for demonstrating purposes, this report uses a very high number of degrees, and eliminates nodes with low degrees. In addition, when we apply different statistics, it will be easier to comprehend the values on a smaller number of nodes.

For this particular filtration, nodes with a degree of more than 32 were selected, therefore all the nodes less than 33 will get eliminated and we are left with Fig.6, where we can clearly see the number of nodes and the edges between them. This is easier to work with in terms of applying statistics and answering questions. Annexed are different statistic measures we will analyze.

A. Degree Distribution

This is the simplest measurement among all the different measurements we will see in this report. The degree distribution helps us observe each node more carefully. There are times when we want to know which nodes share the same degree or which node has a higher degree or a lower degree. In Fig. 7, we can see that Daniel Boulud, Logan Junior, Eric Ripert, and David Chang have the highest degree. We know that we have 11 nodes, and out of 11 we have 4 chefs with the highest degree of 10, which means if we pick a random node, the end-point of this node will most likely have a node with the highest degree.

B. Closeness Centrality

This measurement refers to a node's shortest distance to all the other nodes, and it's a way of understanding the importance of nodes in a network. A lot of times people tend to think that because a node has a high degree, it is the most important node, however, that is not always true. Assume we have a node that is located in between two clusters, in this case, every node would have to go through the middle node to find its shortest distance. Therefore, in cases like these, we can say that the middle node also has high importance. In Fig.8 we can see that there are no separate clusters, which is why the nodes with the highest degree still achieve the most importance. In terms of our food sub-network, we can say that Daniel Boulud, Eric Ripert, and David Chang hold the highest importance level than the other chefs. In other words, these are the celebrity chefs that influence the entire network most quickly.

Fig. 5: Entire community of celebrity chefs

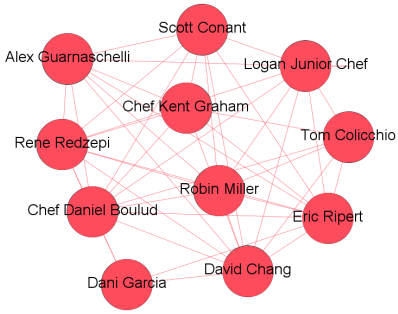

Fig. 6: Filtered community of celebrity chefs

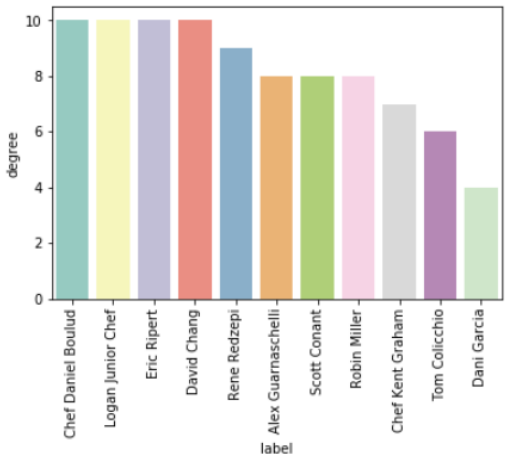

Fig. 7: Degree distribution of celebrity chefs

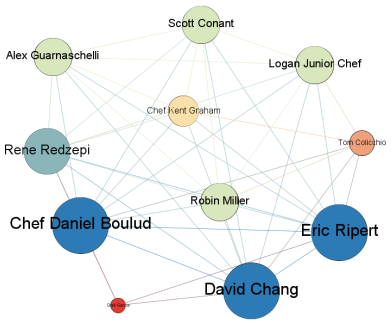

Fig. 8: Closeness centrality of celebrity chefs by size and color (Blue = High, Green = Medium and Red = low)

C. Modularity Class

We have already encountered the use case for this measure, it simply helps us detect community in a network. I thought it would be interesting to see which community can form on a smaller sub-network. When we apply this algorithm, it outputs 2 communities as shown in Fig. 9. In later sections, we will try to research more on these nodes to see why some of these pages form a community.

D. Clustering Coefficient

This measure tells us how well connected the neighboring nodes are of the current node. If the clustering coefficient is 0, that means that there is hardly any connection between the neighboring nodes. If the clustering coefficient is 1, that means all the neighbors of the current node are fully connected with each other. Let's understand this measure using Fig.10, the node with the highest coefficient is shown at the very bottom (Dani Garcia). If we observe all it's 4 neighbors, it shows all of them are connected with each other, hence giving us the coefficient of 1. This measure is useful in our case because we want to know about the people with similar interests.

There are many different statistical measures we can use to analyze our network, however, due to the limitation of this report, only a few important measures were chosen to show. All the addition plots of different measures such as eccentricity,

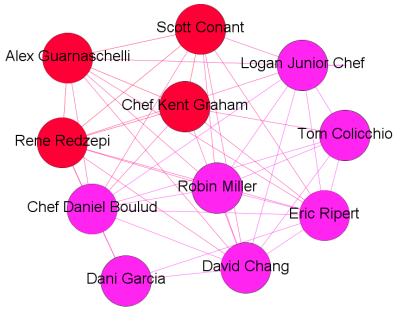

Fig. 9: Community detection on a sub-network

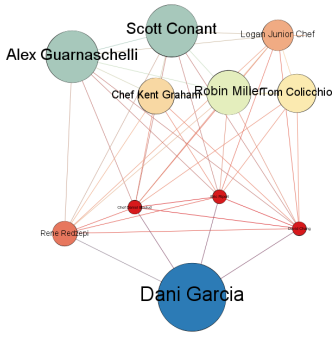

Fig. 10: Clustering coefficient of celebrity chefs by size and color (Blue = High, Green = Medium and Red = low)

closeness centrality, harmonic closeness centrality, betweenness centrality, weighted degree, page ranks, clustering, triangles, and Eigen centrality, are all available in the Jupyter notebook.

V. DRAWING GRAPH BY FORCE DIRECTED PLACEMENT

Imagine we have an integrated circuit (IC) with a bunch of components and wire. How do we connect all the components in such a way that it minimizes the wire length? That is where a placement algorithm comes into the picture. Now, why do we care if a wire is long or short, I mean either way it's going to connect with every component. The answer is distance and energy consumption, the shorter the distance, the faster it is and lower the energy consumption. This is one real-world example, where a force-directed placement algorithm was used for VLSI circuits. Later on, Thomas Fruchterman and Edward Reingold; the authors of this paper [3], based their algorithm on the idea of force-directed placements on undirected graphs. Their algorithm does well at making edge length uniform and symmetric by distributing vertices evenly. According to the paper, one of the major advantages of using this algorithm is speed. It also does a very good job of expanding each node and making the entire graph look aesthetically-pleasing. In network analysis, visualizing a graph is the most important task, as it not only shows structures within the graph, but it also shows different emerging patterns.

In the previous section, we ran different statistical measures on the celebrity chef network. In this section, we will run a force-directed placement algorithm on the restaurant network. In order to generate Fig. 11, first filter the entire food network using partition modularity class filter, and only select restaurant classes. After that, we simply apply the Fruchterman Reingold algorithm to visualize the network. At this point, it is still unclear as to what is going on in the network, therefore, we further narrow the resultant graph by applying degree range filter. For simplicity, we try to work with only a few nodes or pages for better understanding. In the next section, we will see how it makes it easier for us to draw conclusions. While applying degree range filter, I came across a situation where I could clearly see some disconnected graphs in this network, and the reason why this happened is that the degree range has filtered some of the edges, essentially a node that was once connected with an edge, is now unavailable, leaving the graph disconnected. We could potentially lose an important link by applying this filter, which is exactly why it is not recommended to remove nodes from a graph. Even after applying the degree range filter, we can clearly see many visible nodes as shown in Fig.12, to be precise there are 72 nodes and 77 edges. Again, we have to run the modularity measure because the size of our network has changed. The number of communities it formed is 8, with the resolution of 1, and modularity of 0.756.

A. Hits, Hubs, and Authorities

Hubs and authorities are part of HITS or Hyperlink induced topic search, it is a link analysis algorithm that measures the importance of pages. Generally, a good authority is a page that was linked from many good hubs and a good hub is a page that points to many good authorities. One of the interesting observations that were made in Fig.12, is that the sub-network represents both the hub and authority at the same time. For instance, the "Food Panda" restaurant is a really good authority page, and it turns out it's also a good hub. Changing hubs and authority attributes in Gephi did not change the node size at

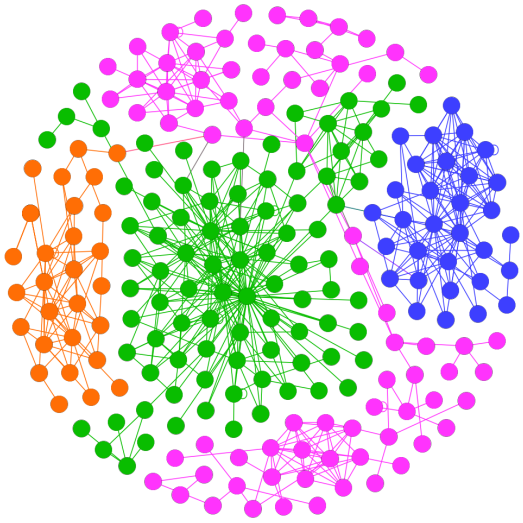

Fig. 11: Forced directed placement on restaurants network

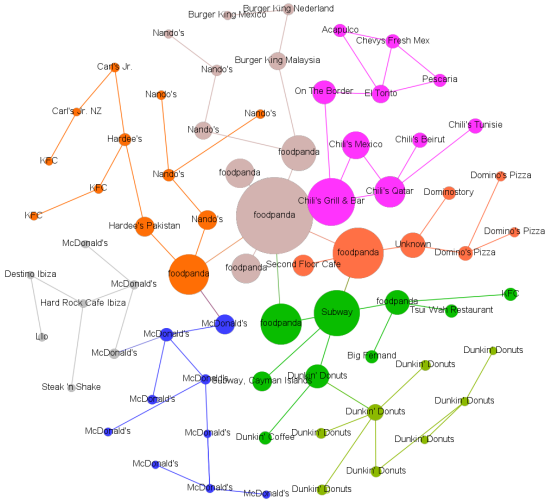

Fig. 12: Filtered community of restaurants with applied authority and hub size

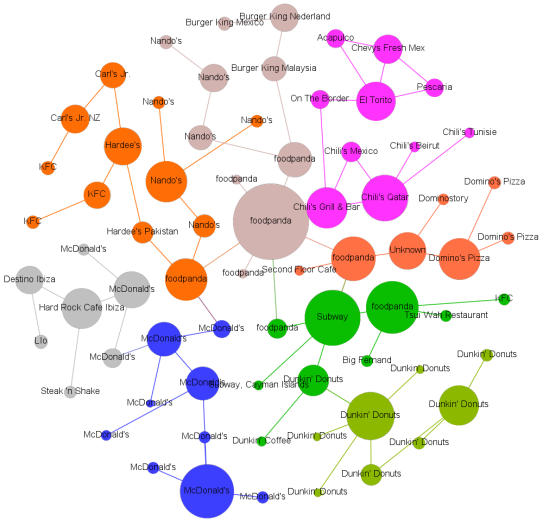

Fig. 13: Community of restaurants with applied PageRank

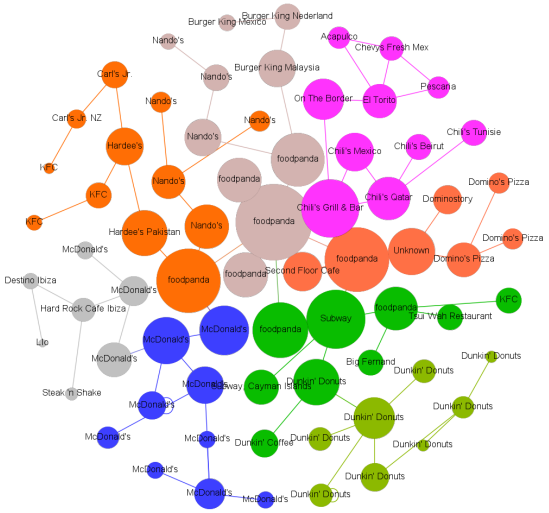

Fig. 14: Community of restaurants with applied harmonic centrality

all, therefore, we can conclude that all pages in this sub-network have similar or close to similar values between the authority and the hub.

B. PageRank

The idea of PageRank is somewhat similar to HITS algorithm, it assigns a score to every node based on their connections, it tells how important a page is, we can think of it as a variant of Eigen centrality. This measure is not of much use for our analysis because it takes into account edge weight and direction. From earlier, we know that our network is undirected with a weight of 1 for every edge. However, it is still used and shown for demonstration purposes. In Fig.13, we can see that "Food Panda" is still ranked the highest among all the other restaurants.

C. Harmonic centrality

This algorithm is a variant of closeness centrality; it is a distance-based centrality measure, and it supports undirected and unweighted graphs. The use cases of this algorithm are similar to that of closeness centrality. Imagine if we want to find important nodes or pages, or even key influencers on a social network. In our case, most of the important pages are located towards the center of the graph as shown in Fig.14.

VI. RESULTS AND DISCUSSION

Before we analyze any results, let's compare the multilevel agglomerative edge bundling method and forced directed placement method. In terms of speed, authors of both the methods concluded that the time complexity is the best for large networks. While I applied the algorithms, I did not notice any significant difference between them. Note that the network we used in this report is small compared to other social networks. Most likely we would find a difference when running these algorithms against a big network. Edge bundling is very useful when your goal is to remove cluttering and want your branches to be separated in a way where it reveals meaningful node clusters. It almost looks like the roots of a tree. Visually, that's a big advantage when it comes to large graphs. Forced directed placement method is very useful when you are trying to analyze a small graph, because everything is condensed in a way where all the nodes are equally spread out. There is no need to scroll around as graphs get very manageable when it comes to applying different statistical measures.

One of the most important tasks in link analysis is to draw different conclusions using metrics. Initially, we divided the entire network into 2 different sub-networks, the reasoning behind this is because the food network was naturally forming 2 different communities. Even though the modularity class detected 5 classes, logically we know that it forms 2 classes, namely celebrity chef network and restaurant network. Let's get started by first analyzing the celebrity chef network.

One of the first measurements we looked at, is the degree distribution. By using this metric we conclude that if a user randomly likes a page of a chef, the chances are they might also like a page from this given set { Daniel Boulud, Logan Junior, Eric Ripert, David Chang }. This information can be used for a recommendation system, where we can recommend pages to users with mutual likes. The next metric is closeness centrality, essentially a node or a page with high closeness centrality, holds the highest importance level, and they acquire vital information and resource within a network. For our network, if we take the top 3 highest closeness centrality values, we get the pages of the following set of people { Daniel Boulud, Eric Ripert, David Chang }, this set perfectly makes sense, because it contains the highest degree among all the other chefs. During my research, I found that Eric and David both acted in a television show called Treme on HBO, which is one possible reason there's a mutual link between the two. The next metric is the modularity, with the help of this metric, we can detect communities. In our case, we have 2 sets of communities. C1 = { Logan Junior, Tom Colicchio, Eric Ripert, Robin Miller, David Chang, Dani Garcia, Daniel Boulud } and C2 = { Scott Conant, Kent Graham, Rene Redzepi, Alex Guarnaschelli }. C1 and C2 were really interesting sets because I researched all of them, and it turns out chefs in C2 are not as popular as C1. Some of the traits C1 have are that they are a TV personality, authors, owner of restaurants, and they have entrepreneur spirit in them. The final statistical measure is the clustering coefficient, we previously learned that its value is high when its neighbor is fully connected with each other. In our example, we found that Dani Garcia has the highest coefficient value, and it is connected with 4 other nodes C3, where C3 = { Rene Rendzepi, David Chang, Daniel Boulud, Eric Ripert }, and they are also connected with each other. Similar to degree distribution, we can use this information to build a recommendation system, where if a person likes Dani Garcia, then they might also like a member contained in C3.

Analyzing the restaurant network is trickier compared to the celebrity chef's network, because most of the nodes in this network are similar in a sense, they have exact labels. I would imagine these restaurant pages are of different locations, but since we don't have a unique location in the dataset, it gets difficult to analyze. However, I found a really interesting pattern, if we carefully observe Fig 12 - Fig 14, we see that Food Panda is always ranked the highest among all the other nodes. After doing more research on Food Panda, it turns out, it is a mobile food delivery service, active mostly in Romania, Bulgaria, and the Asia Pacific. It allows users to select a local restaurant and place orders using their website and mobile app. This would answer why it is always the highest ranked. Most likely people who order food from this service are also getting redirected to their Facebook page for likes.

VII. CONCLUSION

In this report, we analyzed the food pages network which contains mutually liked Facebook pages from the food category. Upon more investigation, the network revealed that there are 5 communities in total, however, in reality, there were only 2 types of community namely chefs and restaurants. Whenever we analyze a network, it is up to us to logically bundle similar classes for further analysis. In order to visualize the chef network, we used the multilevel agglomerative edge bundling method. It helped us alleviate unnecessary cluttering from a network by revealing high-level edge patterns. The second method we used, is called force-directed placement, this method was chosen for the restaurant network because it works really well on small networks. There are no advantages or disadvantages of using either of these methods, the usage highly depends on the type of network and the goal of the project. Visualizing a network is part of the link analysis process, the next step is to calculate different metrics so we can make sense of the network itself. There are many different statistical measures one can apply, however, only a few chosen were discussed in this report. Using those metrics, we drew some conclusions from our network. In an ideal world, we would have a lot more information about the dataset and due to those limitations, only a few conclusions were drawn.

Link analysis helps us find patterns from different relationships and we can discover new knowledge using those relationships, and it is why finding connections is critical. It has been highly used in law enforcement, fraud detection, PageRank, and many more. It helps us explore the relationships between entities that otherwise would be impossible to visualize. In finding all this

data, linking it, and then analyzing it, can help in the fields of business, law enforcement, banking, investing, and medicine, just to name a few. It establishes important connections and is crucial for understanding spheres of influence in a vast array of situations. It can help in determining where businesses need to increase advertising, for them to increase spending in certain areas of their business, or simply to implement new ideas based on the information shown to them through link analysis.

REFERENCES

[1] A. Rossi Ryan, K. Ahmed Nesreen, "The network data repository with interactive graph analytics and visualization", FB Pages Food, 2015, networkrepository.com/fb-pages-food.php.

[2] E. Gansner, Y. Hu, S. North, C. Scheidegger, "Multilevel agglomerative edge bundling for visualizing large graph", 2011, yifanhu.net/PUB/edge_bundling.pdf.

[3] T. Fruchterman, E. Reingold, "Graph drawing by force-directed placement", Vol. 21(11), 1129-1164 (Nov 1991), citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.8444&rep=rep1&type=pdf.