Having +6000 articles sitting on my hard drive, and a few days off from work, I thought it could be interseting to see how the scholars are related to each other in the field of **Syntethic Biology**.

I have done an analysis on **co-authorship** to find the dynamic among scholars, to assess collaboration trends and to identify leading scientists and organizations. The analysis reveals the social structure of the networks by identifying actors and their connections.

To do so, I picked up a random sample of 85 articles with more than one author to creat my network, scraped the authors, perform entity resolution to match similar authors, assign each as a node in the network, and directed edges go from all the authors of one articles to each others. Finally, common network metrics are conducted.

Different centrality measures (explanation in the appendix) for each author has been calculated. These measures include: In Degree Centrality, Out Degree Centrality, Betweenness Centrality, Closness Centrality, Eigenector Centrality, and Bonacich Centrality.

In the table below, the top 30 authors out of 243 are shown in table 1.

	degree	in_degree	out_degree	btwn	close	vector	bon....bonpow.graph.
jay d keasling	31	28	3	88.00000	2.128203e-05	1.000000e+00	6.589077e-03
richmond sarpong	14	13	1	0.00000	2.126664e-05	6.358923e-01	8.236346e-03
nacyra assad garcia	13	13	0	0.00000	1.796622e-05	0.000000e+00	0.000000e+00
yoichiro shiba	18	12	6	59.91667	2.127162e-05	8.113380e-01	-9.733863e-03
radha krishnakumar	13	12	1	0.00000	1.796622e-05	0.000000e+00	1.647269e-03
james kirbya	18	11	7	41.91667	2.127162e-05	8.113380e-01	-5.316070e-03
sydnor t withers	14	11	3	0.00000	2.126664e-05	6.358923e-01	1.003337e-02
sanjay vashee	13	11	2	0.00000	1.796622e-05	0.000000e+00	4.941808e-03
michelle c y chang	19	10	9	47.25000	2.126755e-05	8.132248e-01	-1.512469e-02
chuck merryman	13	10	3	0.00000	1.796622e-05	0.000000e+00	1.153088e-02
monzia m moodie	13	9	4	0.00000	1.796622e-05	0.000000e+00	2.470904e-02
drew endy	10	8	2	23.00000	2.125986e-05	5.907807e-03	6.589077e-03
rachel a eachus	18	8	10	13.91667	2.126709e-05	8.104664e-01	-1.467520e-02
weiss ron	8	8	0	0.00000	1.758396e-05	0.000000e+00	0.000000e+00
timothy s ham	14	8	6	0.00000	2.126664e-05	6.358923e-01	4.567779e-03
li ma	13	8	5	0.00000	1.796622e-05	0.000000e+00	5.106534e-02
paul rabinow	9	7	2	26.00000	2.126348e-05	1.166522e-02	3.294538e-03
smolke christina	8	7	1	0.00000	1.758396e-05	0.000000e+00	1.647269e-03
michael g montague	13	7	6	0.00000	1.796622e-05	0.000000e+00	1.037780e-01
dae kyun robb	22	6	16	201.00000	2.127207e-05	9.676815e-01	-2.343626e-02
michael j selgelid	12	6	6	235.00000	2.127252e-05	2.385723e-01	-1.227964e-02
modrich paul	8	6	2	0.00000	1.758396e-05	0.000000e+00	4.941808e-03
kimberly a ho	14	6	8	0.00000	2.126664e-05	6.358923e-01	-2.245108e-03
gwynedd a benders	13	6	7	0.00000	1.796622e-05	0.000000e+00	2.092032e-01
ian kerridge	5	5	0	0.00000	2.125218e-05	3.576756e-02	-7.315345e-18
keasling jay	8	5	3	0.00000	1.758396e-05	0.000000e+00	1.153088e-02
malcolm campbell	5	5	0	0.00000	1.736232e-05	0.000000e+00	0.000000e+00
john m ndungu	14	5	9	0.00000	2.126664e-05	6.358923e-01	-2.842947e-03
mikkel a algire	13	5	8	0.00000	1.796622e-05	0.000000e+00	4.200536e-01

Correlation between the centrality measures are shown in the table below.

	degree	in_degree	out_degree	btwn	close	vector	bon....bonpow.graph.
degree	1.0000000	0.7859468	0.7318718	0.40243857	0.41615904	0.72964907	0.21093391
in_degree	0.7859468	1.0000000	0.1538806	0.31432660	0.30189626	0.56359515	-0.05026210
out_degree	0.7318718	0.1538806	1.0000000	0.29670324	0.33232966	0.54488791	0.39248684
btwn	0.4024386	0.3143266	0.2967032	1.00000000	0.35272624	0.42692694	-0.02365389
close	0.4161590	0.3018963	0.3323297	0.35272624	1.00000000	0.55107297	-0.01353271
vector	0.7296491	0.5635951	0.5448879	0.42692694	0.55107297	1.00000000	-0.03543419
bon....bonpow.graph.	0.2109339	-0.0502621	0.3924868	-0.02365389	-0.01353271	-0.03543419	1.00000000

A further analysis requires regressing attributes on centrality measures. The following shows the correlation between “in and out degree”, and “total degree and betweeness” by fitting a linear regression model.

```
Call:
lm(formula = in_degree ~ out_degree)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0877 -2.0525 -0.8829  0.9475 25.6083

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.88293    0.25404   7.412  2.1e-12 ***
out_degree   0.16960    0.07015   2.418   0.0164 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.088 on 241 degrees of freedom
Multiple R-squared:  0.02368,   Adjusted R-squared:  0.01963
F-statistic: 5.845 on 1 and 241 DF,  p-value: 0.01636
```

The above result shows that there is positive relation between in degree and out degree. Meaning that for every extra unit in in degree, there will be estimated 2.4 extra out degree unit and it is statistically significant. In the other words, for each paper you write with a colleague, you’d be writing two more with others.

```
Call:
lm(formula = total_degree ~ btwn)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7808  -3.1969  -1.1969   0.8031  19.8440

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.19685    0.27068  15.505  < 2e-16 ***
btwn         0.07908    0.01159   6.825 7.09e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.148 on 241 degrees of freedom
Multiple R-squared:  0.162,     Adjusted R-squared:  0.1585
F-statistic: 46.57 on 1 and 241 DF,  p-value: 7.089e-11
```

The above shows that for every extra unit of writing a paper with other people, your betweeness increases 6.8 units (It is a measure of the extent to which a node is connected to other nodes that are not connected to each other. In the other words the total amount of flow it carries if flow between all other nodes passes it), and it is statistically significant.

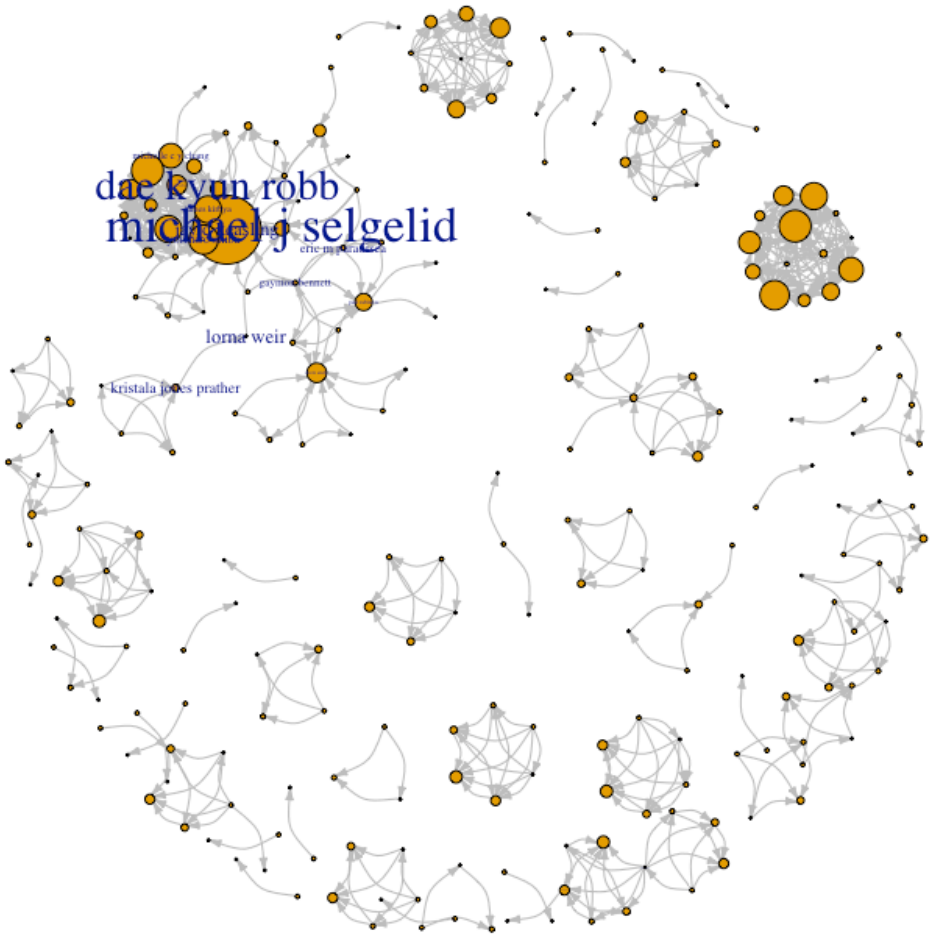
==========

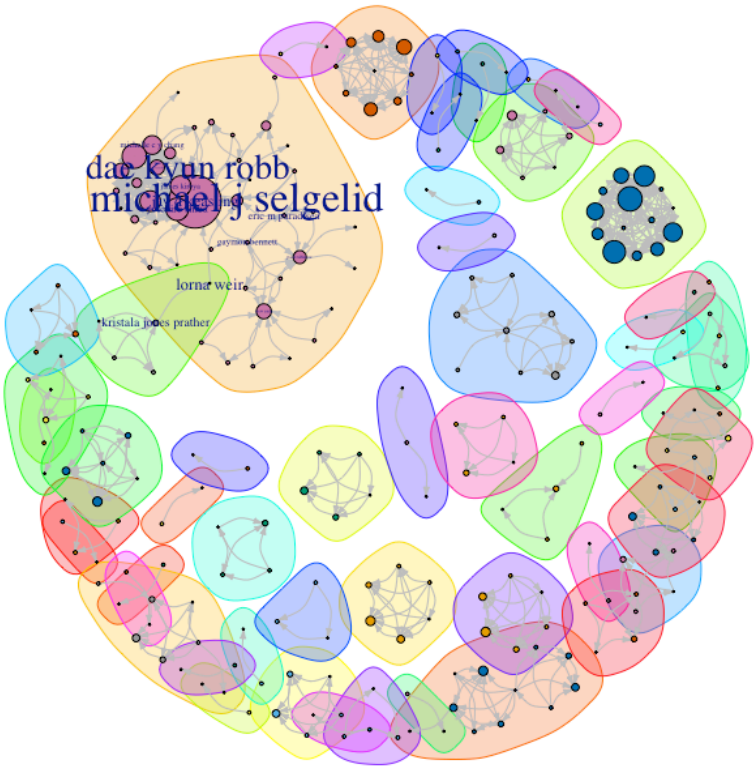
Plotting the graph is another useful analysis. In the following pages, you can see the co-authorship plots with their respective clusters.

The bigger the node names is, the higher is it’s betweenesss.

Also node sizes represent total degree of the node. The bigger it is, the author has done more collaboration with others.

I should remark that there is still a lot to do in the graphics to make it more legible.


dae kyun robb
michael j selgelid
lorna weir
kristala jones prather


dac lynn robb
michael j selgelid
lorna weir
kristala jones prather

Appendix

Measure	Definition	Calculation
total degree centrality	Is the number of ties (edges) a node has without considering their directions. It shows the extend a node holds all of the ties in a network.	The degree centrality of a vertex **v**, for a given graph **G := (V,E)** with **|V|** vertices and **|E|** edges, is calculated as $\mathbf{C_d(v) = deg(v)}$ It is measured by number of edges of a node. It can be also normalized by dividing by total number of edges.
betweenness centrality	Betweenness centrality of node **v** is equal to the number of shortest paths from all vertices to all other that pass **v**. It is a measure of the extent to which a node is connected to other nodes that are not connected to each other. In the other words the total amount of flow it carries if flow between all other nodes passes it.	The betweenness centrality of a vertex **v** for a given connected graph G is: $g(v) = \sum_{s \neq v \neq t} \frac{\sigma st(v)}{\sigma st}$ Where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v. It is between 0 and 1 with 1 being the highest centrality.
closeness centrality	It is based on the length of the average shortest path between a vertex and all vertices in the graph. Is a measure of the degree to which an individual is near all other individuals in a network.	$C_c(i) = \left[\sum_{j=1}^{N} d(i,j)\right]^{-1}$

They all measure the consequences of having (or not having) a certain node in a graph which shows the importance/power/influence of the node in the network structure.

High Degree Centrality:

A node with high degree centrality, has more autonomy from its alters (comparing to a node with lower degree), therefore, it has more flexibility/power/opportunities/choices to pick any of those nodes in for example a trade context.

High Betweenness Centrality:

A node with high betweenness centrality has a large influence on the transfer of items through the network, under the assumption that item transfer follows the shortest paths. Betweenness Centrality is related to connectivity in the graph, in so much as high betweenness vertices have the potential to disconnect graphs if removed. Therefore, a node with high betweenness centrality has the capacity to broker contacts among other nodes, and isolate other nodes or prevent contact.

High Closeness Centrality:

A node with high acts as a reference point by which can reach a larger number of other nodes at shorter path lengths. It emphasizes the distance of an actor to all others in the network (as oppose to direct ties in degree centrality).

Potential Limitation:

Degree centrality is particularly useful when we want to compare networks cohesiveness as in how the ties have been distributed. It addresses the problem that can rise when graph density is not adequate measure (for example if in a network certain node has disproportionately high degree).

However, It's main limitation is that it only takes into account the immediate ties that a node has to its egos, rather than indirect ties to all other nodes. A node with a high total degree centrality might be connected to some whole other nodes, but those nodes might be rather disconnected from the network as a whole, thus the ego node is only central locally and therefore doesn't have the ability to broker between groups (or information/flow originated in other parts of the network is likely not to reach it).

Betweennesss centrality is a good measure to find nodes which bridge subgroups in a network, and consequently indicate structural holes in a network. It can also identify levels

of hierarchy in a network. If one eliminates all the actors with no betweenness (that is, the "subordinates"), some of the remaining actors will then have 0 betweenness—they are at the second level of the hierarchy.

While it is useful for analyses of spread of disease and pandemics, it can be misleading for simpler analysis.

As **closeness Centrality** relies on the sum of the geodesic distances form each actor to all the others, it can be misleading in complicated graphs. A node that is very close to a relatively small subset of a network, can have the same score as a moderately close node to every actor in a large subset of the network.

I rather pick my favorite measure based on the analysis that I am doing instead of having a universal one. As mentioned above, for picking the most popular node, I will use degree centrality. To see which node is passing the most flow I will use betweenness centrality. For analysis similar to market/resource access, I will use closeness centrality.