課　程：網路科學

教　授：謝宏昀 老師

題　目：Homework Assignment #1

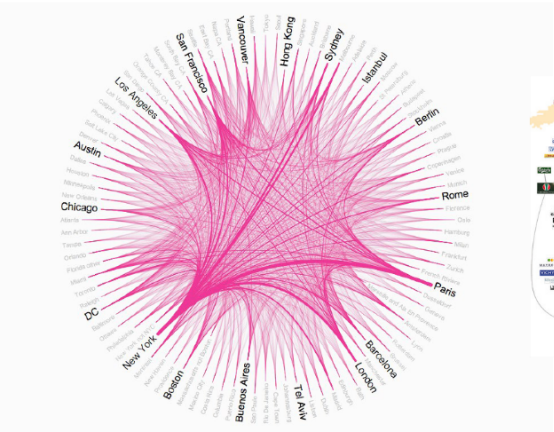
學　生： 陳璽文

學　號： B02901017

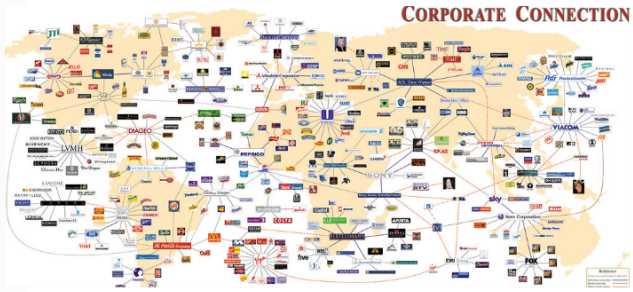
日　期： 20160404

github : link

1. (20%) Surf VisualComplexity.com to find out what the website is about. Pick two projects from two different categories that interest you most. Describe the formation of networks (e.g. what do nodes and links represent, are links directed or weighted?) and what the network is used for. Explain what interests you most about the networks.



(a) top 50th market of Airbnb network



(b)300 brand in US network

A. Case I : Airbnb link

The right graph represent the airbnb service usage in the top 50th markets all over the world. Each nodes represent a cities, with the thickness of the lines corresponding to the relative volume of travels between each pair. The Reason why they use the circle layout is that there are few nodes, and the transactions between each cities are more like pair to pair rather than a serial of nodes. In this case, we can found out that the most Asia cities have less willingness to share there house except Hong Kong. Besides, the popular tourist attraction have stronger edge.

B. Case II : Corporate Connection link

tis network describe the connection between 3 celebrities, 35 corporations, 40 subsidiaries and more than 300 brands. It’s an undirected graph. It’s really obvious that the it form a close cooperation between each brands. It draw as a spring layout because there is several local center in this network, which is the cooperations . Besides, there is still some interesting issue. For example, this network can separate to several communities through industries. And the closest community to the media is the food industry. It’s quite similar to Taiwan.

2. (60%) Download the Amazon product co-purchase dataset and first use any text editor to open it. Each line of the dataset represents one undirected link between two products indexed by their product IDs, meaning that the two products are frequently co-purchased based on the user purchase history maintained by Amazon. To probe for useful information from the dataset, open the dataset using any network analysis and/or visualization tools of your choice. Note that the dataset contains a total of 334,863 nodes and 925,872 links, so you need to choose a tool that can handle such a network size.

   A. Provide statistics to describe the entire network, including the degree distribution (together with mean, min, max), network density, diameter, mean geodesic distance, clustering coefficient, assortativity coefficient (degree correlation coefficient), and so on. What do these numbers tell us about the product co-purchase network?

   Form our data set we can summarize as the following form:

		All		Book		Music		DVD	
		value	time(s)	value	time(s)	value	time(s)	value	time(s)
#nodes		334863	7.8e-06	126219	5.0e-06	27676	6.9e-06	5891	5.0e-06
#edges		925872	1.001	180283	0.0744	35549	0.0147	5263	0.005
network density		1.65e-05	0.995	2.263e-05	0.0752	9.283e-05	0.0155	3.1e-04	0.0039
assortativity		-0.0588	4.350	-0.069	1.0131	0.0385	0.2112	0.0203	0.0345
diameter		X	X	X	X	infinite	0.304	infinite	0.229
mean geodesic distance		X	X	X	X	infinite	0.316	infinite	0.226
Degree	max	549	0.987	123	1.272	46	0.0165	34	0.0047
	min	1		0		0		0	
	mean	5.53		2.857		2.568		1.789	
Betweenness	max	X	X	X	X	0.0531	2621.42	0.0115	71.89
	min	X		X		0.0		0.0	
	mean	X		X		0.0002		0.0001	
Closeness	max	X	X	X	X	0.0467	286.79	0.0124	0.860
	min	X		X	X	0.0		0.0	
	mean	X		X	X	0.0159		0.0018	
eigenvector	max	error	error	0.4918	17.71	0.4497	8.916	0.4567	0.382
	min	error		0.0		0.0		0.0	
	mean	error		0.0001		0.0003		0.0013	
clustering	max	1.0	6.571	1.0	1.077	1.0	0.171	1.0	0.027
	min	0.0		0.0		0.0		0.0	
	mean	0.396		0.277		0.3058		0.1984	

The ‘X’ means that it takes more than 3 hours. As the function have implement the shortest path Alg. The time complexity is in $O(V^2)$, which is really hard to calculate. As a result I cut the

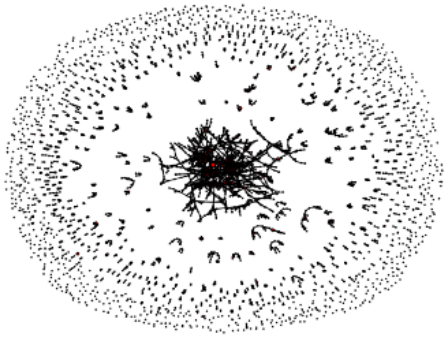
graph through the meta data in to three categories : Book, DVD and Music. The main different is that after cutting the graph become a none connected graph. As a result, there not only might be some important nodes to connect each categories In this network, but also some node connect between two products in same category. Another interesting thing is that even though the complexity of closeness is $O(v^2)$, this object have save the shortest path matrix and inherent by other function. The time performance is better than when we first calculate it.

Besides, after cutting, the assortativity change the sign after we build the three subgraphs. It seems that there might be several edge between the popular node in each categories link to not frequently purchasing node in other categories. However, in one categories the central nodes tend to link each others .

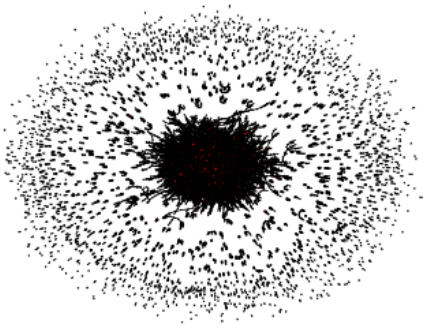
B. Identify "special" nodes based on different node-based measures introduced in class including the centrality measures and local clustering coefficient. What do these numbers tell us from different perspectives about the specialty of these products? Download the Amazon product metadata file regarding the names of these products and provide your reasoning why the products identified in (b) are so special?

According to the above form we can got the following picture, we plot the points through the spring layout. The points size depend on the value with the factor of average, so the bigger the point is, the higher the value.

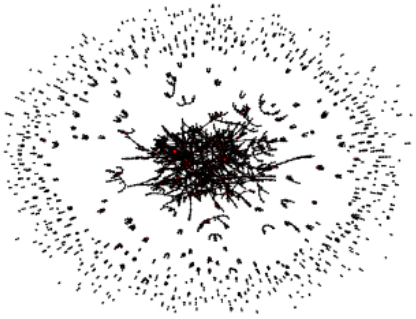
(a) top 5000th degree of all network



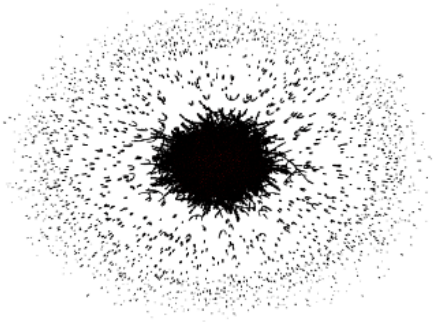
(b) top 5000th degree of book network



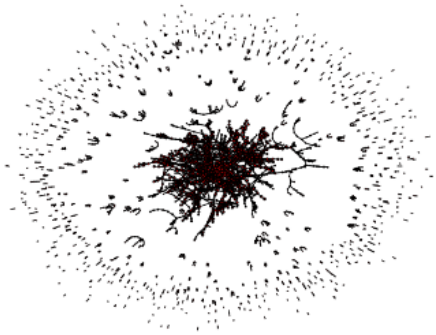
(c) top 5000th degree of book Music network



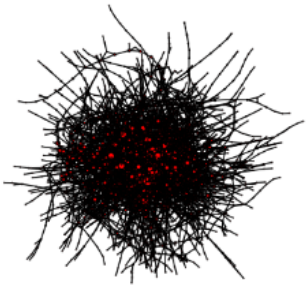
(d) top 5000th degree of book DVD network



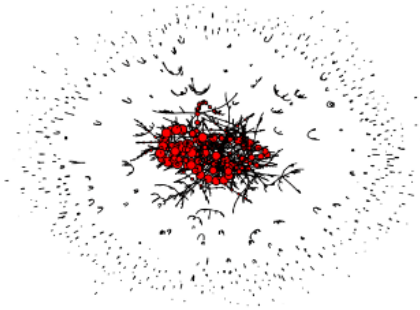
(e) top 5000th closeness of book Music network



(f) top 5000th closeness of book DVD network



(g) top 5000th betweenness of book Music network



(h) top 5000th betweenness of book DVD network

We can find out that the all network become more sparse than three of others. And its similar to the Book graph. It’ s quite trivial that the main characteristic is dominant by the book. And the sub-network have better connectivity.

Than we focus on the difference sub-network : Music and DVD. According to the picture we can notice that the Music have the better connectivity in the central node. The even stick together become a K-clique graph. On the other hands DVD have the better local centrality. The central node might link to several unimportant nodes, just like a hub. Than each hub connect to each others.

The following form is that we list the most central mode from different measurement:

		All	Book	Music	DVD
degree	ID	548091	199628	186810	20594
	Name	Laura (DVD)	The Great Gatsby	Dark Side of the Moon 30th Anniversary Edition	Hellsing - Complete Collection
	value	549	123	46	34
betweenness	ID	X	X	259613	236257
	Name	X	X	Ray Charles: Ultimate Hits Collection	A Prayer For The Dying
	value	X	X	0.0531	0.0115
closeness	ID	X	X	259613	20594
	Name	X	X	Ray Charles: Ultimate Hits Collection	Hellsing - Complete Collection

	value	X	X	0.0467	0.0124
eigenvector	ID	X	515301	301097	44918
	Name	X	1001 Most Useful Spanish Words (Beginners' Guides)	The Very Best of the Spinners [Rhino]	Hellsing - Complete Collection
	value	X	0.4918	0.4497	0.4567

We can also found out that there are several repeat products in this form. It’s quite reasonable because the centrality might have some similarity. And those product have high sales rank, and it might frequently vanish in the similar tag.

C. Choose a product and calculate its similarity with other products based on the two similarity measures introduced in class. Find the “most similar” products of the chosen product. How do you interpret the results based on these similarity measures? Does it agree with our common sense? (Find out product names from the meta- data file.)

We choose the top 10 similar pair from the DVD sub-network and fill the following form:

rank	ID1	name1	ID2	name2	rate
1	16650	In Too Deep	272315	Out of Sync	0.848
2	59463	The Thin Man	74222	Arsenic and Old Lace	0.841
3	249822	The Legend of Drunken Master	6292	Winners and Sinners	0.838
4	79816	Musik Trienniale Koln 2000 - Berg Lulu Suite / Debussy Le Jet D'Eau / Stravinsky Firebird / Boulez, Chicago Symphony Orchestra	79815	Stravinsky - Le Sacre du Printemps (The Rite of Spring) / Symphonies D'Instrument Vent / Boulez, London	0.836
5	119344	Leave It to Beaver	30810	Mouse Hunt	0.833

We can notice that the no.1 are both gangster films, no.2 are both 40th American old movies. The no.3 are both 90th cinema of Hong Kong, which are starred by Jackychen. No.4 are both Symphony Orchestra. And the no.5 have the less similarity one is a 60th TV show the other is a 90th movie, however there are both comedy.

Interestingly, four of our measurement is in the similar row in the meta data.

D. Amazon plans to re-design its product main page and select several products to highlight. What products do you suggest Amazon to put on the main page based on what you have learned from the dataset? Explain your reasoning. Do you have any other suggestion for Amazon based on - your crawling of the dataset?

I would suggest a recommended system depends on the Katz centrality, because it can represent not only the importance of the itself, but also concern about neighborhood contribution. And the algorithm might be a weighted graph that the weight is relative to the log time . The newer the order is , the higher the weight is. As a result we can catch the current trend rather than promote the same products again and again.

3. (20%) Parse the metadata file of the Amazon product dataset mentioned in 2(b). Associate one feature (or, characteristic) with each product (e.g. sales rank, product category, average rating). Does the network exhibit assortative mixing for the chosen feature? Find out relevant measures and provide your reasoning.

   First each central nodes have the high sales rank. Most of them are less then one thousand. Among four measurements, the degree has the highest correlation, and the betweenness are the least. It’ s reasonable to say that the def betweenness is like th neck of our network , so it might not be the most popular product, but it connect several subgraph. Second the are all frequently list in the similar rows. And the average rating are all higher than 4.5, more over the votes of the comments are higher than others.