Why Anchorage is not (that) important: Binary ties and Sample selection

http://toreopsahl.com/2011/08/12/why-anchorage-is-not-that-important-binary-ties-and-sample-selection/

Tore Opsahl

See also: Opsahl, T., Agneessens, F., Skvoretz, J., 2010. Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks 32 (3), 245-251.

August 12, 2011 at 1:39 am *4 comments*

A surprising finding when analysing airport networks is the importance of Anchorage airport in Alaska. In fact, it is the most central airport in the network when applying betweenness! Betweenness for a node is defined as the number of shortest paths among all others that passes through the node (see Opsahl et al., 2010, for a review). A host of explanations have been offered to account for the high betweenness of Anchorage given its relatively few connections. For example, Guimera et al (2004, pg. 7797) reasoned that:

Alaska is a sparsely populated, isolated region with a disproportionately large, for its population size, number of airports. Most Alaskan airports have connections only to other Alaskan airports. This fact makes sense geographically. However, distance-wise, it also would make sense for some Alaskan airports to be connected to airports in Canada's Northern Territories. These connections are, however, absent. Instead, a few Alaskan airports, singularly Anchorage, are connected to the continental U.S. The reason is clear: the Alaskan population needs to be connected to the political centers, which are located in the continental U.S., whereas there are political constraints making it difficult to have connections to cities in Canada, even to ones that are close geographically ([Guimera and Amaral, 2004]). It is now obvious why Anchorage's centrality is so large. Indeed, the existence of nodes with anomalous centrality is related to the existence of regions with a high density of airports but few connections to the outside. The degree-betweenness anomaly is therefore ultimately related to the existence of communities in the network.

While many researchers and practitioners highlight this finding, I do not believe it is completely accurate. There are two reasons for this:

Issue 1: Binary ties

Admittedly this might be a personal bias as most of my work has been on weighted networks. Without going into much detail in this blog post, I actually strongly believe that if you assign the same importance to the connection between London Heathrow and New York's JFK as you do to the connection between Pack Creek Airport and Sitka Harbor Sea Plane Base in Alaska (map), then there is a potential for measurement error. The table below lists the top ten airports in terms of betweenness when analyzing the binary and weighted (by passengers) versions of the Bureau of Transportation Statistics (BTS) Transtats data (Brandes, 2001). The code to replicate these results can be found at the end of this page.

	Betweenness	
Rank		
	Binary Analysis	**Weighted Analysis**

	Airport	Score	Airport	Score
1	ANC (Anchorage, AK, USA)	465272	SEA (Seattle/Tacoma, WA, USA)	834217
2	FAI (Fairbanks, AK, USA)	215503	ANC (Anchorage, AK, USA)	761834
3	YYZ (Toronto, Canada, Canada)	131562	ATL (Atlanta, GA, USA)	735628
4	LAX (Los Angeles, CA, USA)	129246	LAX (Los Angeles, CA, USA)	531980
5	SEA (Seattle/Tacoma, WA, USA)	125151	ORD (Chicago, IL, USA)	409001
6	JFK (New York, NY, USA)	124927	DEN (Denver, CO, USA)	314764
7	HPN (White Plains, NY, USA)	121096	JFK (New York, NY, USA)	247791
8	MIA (Miami, FL, USA)	120643	MIA (Miami, FL, USA)	206547
9	DEN (Denver, CO, USA)	120342	BOS (Boston, MA, USA)	168140
10	MSP (Minneapolis, MN, USA)	111188	FAI (Fairbanks, AK, USA)	157491

This table demonstrates that Anchorage has twice the betweenness of the runner-up, Fairbanks Alaska, in the binary analysis. In the weighted analysis, Anchorage loses the first place to Seattle and Fairbanks moves to 10th place. It is also worth noticing that only US airports are in the top ten lists using both analyses, which leads me on to the second issues with using the BTS data: sample selection.

Issue 2: Sample selection

This issue affects all network studies, and something I have been interested in for a while. We define a population and analyse the connections among them. For example, I have analysed the scientific collaboration network based on the papers uploaded to the arXiv preprint server (e.g., Opsahl et al., 2008) with the full knowledge that there are many more scientific publications out there as well as other forms of collaboration and channels for knowledge flow among scientists, such as grant proposals and conference attendance. By simply restricting ourselves to data that is easy to collect (often stored in a central location / repository), the research is vulnerable to sample selection bias.

When it comes to airport networks, the Bureau of Transportation Statistics (BTS) Transtats data is straight forward to collect: Go here, select what you want (Origin, Destination), and click Download! However, there is a small note on another page explaining the dataset: "*This table combines domestic and international market data reported by U.S. and foreign air carriers, and contains market data by carrier, origin and destination, and service class for enplaned passengers, freight, and mail. For a uniform end date for the combined databases, the last 3 months U.S. carrier domestic data released in T-100 Domestic Market (U.S. Carriers Only) are not included. Flights with both origin and destination in a foreign country are not included.*" It is the last line of this description that highlights the potential sample selection bias. While the data contain all US airports and all domestic flights, it only contains Non-US flights that leave or terminate at a US airport and the Non-US airports on the other end of these flights. As such, a section of the square adjacency matrix is missing (flights from Non-US to Non-US airports in the dataset) as well as the entire rows and columns for airports without flights to the US. To exemplify this bias, I have plotted the routes on a world map below.

The BTS airport network in 2010 where the line colour is based on the number of passengers. The code to replicate this image can be found at the end of this page. If you click on the image, a vector graphic version of it is available (pdf; 5.95mb).

While it is possible to see a concentration in the US on the above picture, the sample selection becomes much more apparent when highlighting Europe. In the picture to below, it is possible to see that no flights are between any pair of European cities nor any other point on the map. Would you have to transit at New York's JFK to get from London to Barcelona? This gap highlights the need for looking for more complete data sources than the Bureau of Transportation Statistics when analysing airport networks.

The European part of the BTS airport network

Alternative data-source

There are a couple of authoritative databases with world-wide airline routes. However, most of them are proprietary as they have enormous business intelligence potential and, as a consequence, are difficult to collect. OAG Worldwide is one such database, and it should be noted that Guimera et al (2004) went through the hoops by getting this data, and therefore, had a much more complete view of the airport network than if they had used the BTS Transtats data. While I do not have access to such a database, Openflights.org is a crowdsourced alternative. Although using this data comes without any guarantee, it has the potential to showcase the limitations of the BTS Transtats data. As a first step, I mapped the data to ensure there were no obvious pockets of missing data.

The Openflight.org airport network where the line colour is based on the number of routes (accessed on August 12, 2011). The code to replicate this image can be found at the end of this page. If you click on the image, a vector graphic version of it is available (pdf; 5.25mb).

Conclusion 1: Anchorage is not the most important airport

As can be seen from this picture, there are no obvious areas without any form of airline traffic. To show how this data impacts on a betweenness analysis, I have computed betweenness on both the binary and weighted (by number of routes as the passenger numbers were not available) versions of the network. As can be seen in the table below, major airports located around the globe get the highest scores in these analyses instead of only US airports. Specifically, Anchorage is only the third most central in the binary analysis, and the 14th most central in the weighted analysis. As such, it is still an important airport in the networks, but maybe not the most important.

	Betweenness			
Rank	**Binary Analysis**		**Weighted Analysis**	
	Airport	**Score**	**Airport**	**Score**

1	FRA (Frankfurt, Germany)	587531	LHR (London, United Kingdom)	1858349
2	CDG (Paris, France)	520707	LAX (Los Angeles, United States)	1310287
3	ANC (Anchorage, United States)	481044	JFK (New York, United States)	1084392
4	DXB (Dubai, United Arab Emirates)	443314	BKK (Bangkok, Thailand)	797785
5	GRU (Sao Paulo, Brazil)	402882	SIN (Singapore)	739981
6	YYZ (Toronto, Canada)	398869	SEA (Seattle, United States)	723145
7	LHR (London, United Kingdom)	389846	MAD (Madrid, Spain)	707354
8	LAX (Los Angeles, United States)	356600	GRU (Sao Paulo, Brazil)	684057
9	DME (Moscow, Russia)	353902	NRT (Tokyo, Japan)	639074
10	BKK (Bangkok, Thailand)	352682	DXB (Dubai, United Arab Emirates)	610765
…	…	…	…	…
14	…	…	ANC (Anchorage, United States)	469203
18	…	…	FRA (Frankfurt, Germany)	392418

_

Conclusion 2: Finding the global superhub using a weighted approached

London Heathrow is the most central airport when considering both tie weights and the global airport network. And this, unlike Anchorage, is not a surprising finding as it is the airport with most international passengers (Airports Council International, 2011).

To further investigate the effects on the ranking when considering tie weights in the global airport networks, I considered the change in ranking of the two airports ranked first in the binary and weighted analyses, Frankfurt and London Heathrow. Frankfurt went from having the highest betweenness in the binary analysis to only having 18th highest betweenness in the weighted analysis. Conversely, London Heathrow went from having the seventh highest to the highest betweenness score. To look into this cross-over of rankings, I compared the degree (number of airports with direct flights) and strength (number of direct routes) from these two airports:

			Strength distribution				
Airport	**Degree**	**Node Strength**	**1**	**2**	**3**	**4**	**5**
FRA (Frankfurt, Germany)	237	349	142	82	9	4	0
LHR (London, United Kingdom)	157	288	71	55	22	4	5

This table shows that Frankfurt has direct flights to 51% more airports than London Heathrow, but only 21% more routes. The variation in tie weights can be further investigated by looking at the weight distribution. While there are only four airports with four direct routes from Frankfurt, there are nine airports with four or five direct routes from London Heathrow.

Moreover, by looking at which airports have the strong ties (i.e., with tie weights greater or equal to 4) with Frankfurt and London Heathrow, it is possible to see that the geographical distribution is strikingly different. Frankfurt has four direct routes to Antalya (Turkey), Madrid (Spain), Mallorca (Spain), and Vienna (Austria), which are 5,597 kilometres long (average: 1,399km). Conversely, London Heathrow has five routes to Delhi (India), Dubai (UAE), Hong Kong (China), Los Angeles (LAX, USA), and New York City (JFK, USA) and four routes to Bangkok (Thailand), Mumbai (India), Boston (USA), and Miami (USA), which are 65,376 kilometres long (average 7,264km). By having strong ties to geographically distant instead of close airports, London Heathrow acts as a intercontinental hub instead of a continental hub. Additionally, the airports with strong ties to London Heathrow have high betweenness, and therfore, act as hubs in their respective regions. As such, London Heathrow can be seen as the global hub of the world-wide airport network.

References

Airports Council International, 2011. Year to date International Passenger Traffic, Apr-2011, accessed August 12, 2011.

Brandes, U., 2001. A Faster Algorithm for Betweenness Centrality. Journal of Mathematical Sociology 25, 163-177.

Bureau of Transportation Statistics, 2011. Air Carrier Statistics (Form 41 Traffic): T-100 Market (All Carriers), accessed August 12, 2011.

Guimera, R., Amaral, L. A. N., 2004. Modeling the world-wide airport network. The European Physical Journal B 38, 381–385.

Guimera, R., Mossa, S., Turtschi, A., Amaral, L. A. N., 2004. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. Proceedings of the National Academy of Sciences 102(22), 7794-7799.

Openflights.org, 2011. Airport, airline and route data, accessed August 12, 2011.

Opsahl, T., Agneessens, F., Skvoretz, J., 2010. Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks 32 (3), 245-251.

Opsahl, T., Colizza, V., Panzarasa, P., Ramasco, J. J., 2008. Prominence and control: The weighted rich-club effect. Physical Review Letters 101 (168702).