The Statistical Mechanics of Scale-Free Networks

Wade DeGottardi

December 10, 2007

Abstract

The methods and ideas from the emerging field of scale-free networks have been applied to a diverse group of problems. In this paper we survey some important theoretical developments and look at several prominent studies. Possible future directions in the field are discussed.

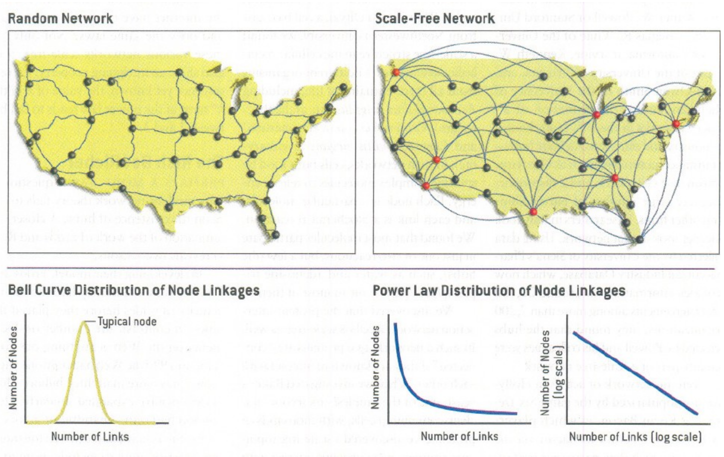

Figure 1: Random and scale-free networks (from [2]).

1 Introduction

The application of statistical mechanics to networks has proven to be relevant to a wide variety of social, biological and technological networks. Unlike many areas of physics which focus on microscopic or local interactions, the study of networks employs inherently *inductive* techniques. Systems are studied using various metrics and then plausible models are built which attempt to make contact with empirical studies. This approach is necessary because as pointed out in [1], many network problems have local dynamics which are unknown or ambiguous. Indeed, one of the goals of the fields is to help explain how networks evolve. We begin by looking at the key ideas used in network studies. We then look at several different systems to which these methods have been applied. In this way, we hope to learn the some of the tools of the trade and also look to gain some intuition about networks in general. We finish by discussing some open questions and possible future directions.

2 Network architecture

We begin by considering the simplest possible random graph. Given N nodes we cycle through each possible pair and connect them with a probability p; each connection is called an edge. The resulting distribution of edges per node is bell-shaped and sharply peaked about its mean (figure 2). None of the real-world networks we will investigate in this paper are random graphs. As pointed out in [2], this is not surprising since most networks do not develop in such an artificial way. In fact, we will see that the distribution of edges per node is well-described by a power law in many different kinds of networks (the number of edges per node is called the *degree* of that node). The correctly normalized power law distribution is then given by

$$P(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}. \tag{1}$$

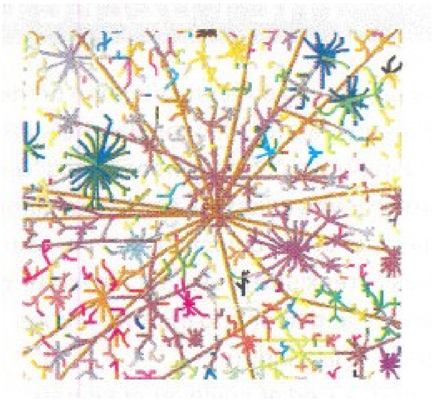

Figure 2: An illustration showing the scale-free nature of the internet. Web pages with similar addresses are connected by similar colors. Note the existence of hubs which connect a large number of sites. A feature of scale-free networks is that they are self-similar: there are large hubs but also smaller hubs, which, in this illustration resemble kaleidoscope eyes. From [2].

Networks that have this distribution are known as scale-free. [2] makes the observation that while random networks resemble highway maps, scale-free networks look more like airline service routes (see figure 1 and figure 2). Unlike random graphs, scale-free graphs have a few nodes of very large degree (hubs). In many social networks a relevant quantity is the average distance between any two nodes. A random graph with N nodes has an average distance that scales as $\ln N$ [1]. Not surprisingly, scale-free graphs are smaller (and in fact are maximally small). For example, the average distance on scale-free graphs with $2 < \gamma < 3$ goes like $\ln \ln N$ [3]. Returning to the map example, if I wanted to travel from Champaign to Santa Barbara, my itinerary would probably take me from Willard to LAX via O'Hare followed by a commuter flight to Santa Barbara. A road trip would take us through many of the major cities between Illinois and California: St. Louis, Kansas City, Santa Fe, etc. Not surprisingly, the existence of hubs in scale-free networks informs many of their salient features.

2.1 Scale-free networks

Why are power laws so ubiquitous in many different types of networks? (see table 1) One of the earliest explanations came from Barabasi and Albert [1]. The Barabasi-Albert (BA) algorithm generates a scale-free network based on the idea of *preferential attachment*. At each time step a node is added to the network and a fixed number of edges are attached to existing nodes. The probability of attachment is proportional to the degree of the target node, a process in which 'the rich get richer' [2]. It's important to distinguish between BA generated networks and the much larger class of random scale-free networks. A random scale-free network is a network in which the distribution of node degree is constrained to have the form of equation 1 but the graph is random in all other respects. The mechanism of preferential attachment is certainly plausible since in many social networks prestige is primarily based on the number of connections that one has made.

There is strong evidence that the web of infection for many diseases also shows scale-free behavior. Most ominously, standard epidemiological models applied to graphs show that for any nonzero transmission rate an epidemic is inevitable on a scale free graph with $\gamma \leq 3$ [4]. This result is related to the fact the variance of the degree of each node diverges in this regime. More generally, it's clear that the ultra-small character of scale-free networks makes a disease easier to spread.

Network	Size	$< k >$	γ
WWW	4×10^7	7	2.38/2.1
Movie actors	2×10^5	28.8	2.3
Co-authors, SPIRES	$56,627$	173	1.2
Sexual contacts	2810		3.4
Metabolic, *E. coli*	778	7.4	2.2
Words, synonyms	22311	13.48	2.8
Avian flu network	3346		1.2

Table 1: Several real-life networks taken from [1]. $< k >$ is the average degree of nodes in the network. γ is the exponent in equation 1. Note that there are two γ's for the WWW. The first describes the degree of incoming links, the second describes outgoing links.

2.2 Building networks

One of the challenges in studying networks is that the topology of a network may not be directly accessible. This is certainly the case for many disease networks. The methods of data analysis presented in [5] are interesting because the network of disease transmission was established indirectly. The study used all known cases of avian flu observed in either wild or domestic birds between November 2003 and March 2007. Specifically, the data consisted of 3346 triples (t_n, λ_n, ϕ_n) where t_n is the time in days since November 25, 2003 of the nth incident; λ_n and ϕ_n are the latitude and longitude of the reported case. Small et al. then constructed a network linking nodes i (t_i, λ_i, ϕ_i) to node j (t_j, λ_j, ϕ_j) if

$$d(i,j) \leq (t_j - t_i)\mu \tag{2}$$

and

$$0 \leq (t_j - t_i) < T_{max} \tag{3}$$

where $d(i,j)$ is the great circle distance between the nodes and μ is a positive constant that represents the maximum rate of transmission of the virus. The raw data and the constructed network are shown in figure 3. In order to test the robustness of this approach, the authors examined various subsets of the data and found that the resulting exponent ($\gamma \approx 1.2$) is not sensitive to different partitions of the data. The fact that this disease network has a scale-free topology has important implications for containment strategies. As argued in [5], identifying hubs in a disease network is an important step in developing immunization policy.

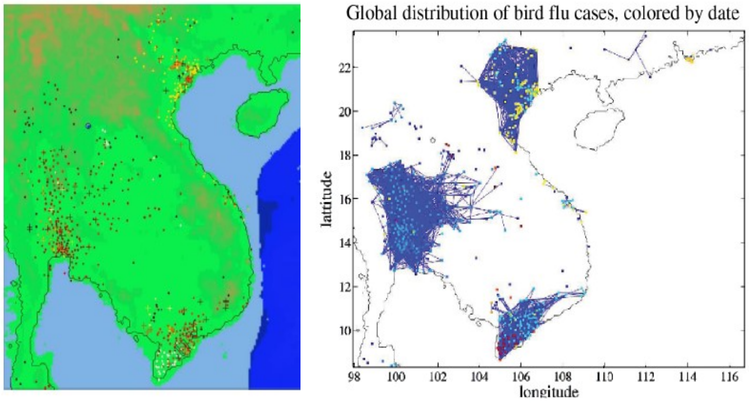

Figure 3: The map on the left shows the locations of avian flu outbreaks. The map on the right shows the constructed network using the conditions in equations (2) and (3). Both graphs are taken from [5].

The idea of nodes connected by edges is a powerful conceptual tool for many problems, even for some where the relevance of the network concept might at first seem suspect. For example, networks have been used to model protein assembly in cells [1]. In order for proteins to function properly, they must adopt a specific conformation [10]. Consider modeling the process of protein assembly by letting nodes represent the different conformations of the protein. Two nodes are then linked if one conformation can be obtained from the other by an elementary step (for example, a fold).

The translation of metabolism into graph theoretic language is more straightforward: biological compounds are represented by nodes. Two nodes are connected if the corresponding compounds participate in the same biochemical reaction. The data used in [6] were taken from the WIT database (http://igweb.integratedgenomics.com) and at the time of the study the database contained information about the metabolic pathways of 6 archaea, 32 bacteria, and 5 eukaryotes. Because metabolism proceeds in one direction, it's possible to get information about the number of incoming and outgoing links. In [6] they found that both incoming and outgoing links were universally scale free with $\gamma \approx 2.2$.

2.3 Hierarchical Networks

Although the BA algorithm generates self-similar networks, a scale-free node distribution alone does not preclude a hierarchical topology. As Ravasz et al. point out [7], while some of the features of metabolic networks seem well described by a scale-free architecture (a few highly connected nodes for example; they point out for example that pyruvate and coenzyme A are ubiquitous in metabolic pathways), there are many other features which are not accounted for by a BA generated network. For example, metabolic networks are known to be hierarchical, a feature not compatible with a purely scale-free topology. An additional metric is clearly needed in order to probe metabolic networks. Ravasz et al. therefore investigated the so-called *clustering coefficient* to further analyze their data. For a given node, the clustering coefficient C_i is given by

$$C_i = \frac{2n}{k_i(k_i - 1)} \tag{4}$$

where n is the number of links connecting the nearest neighbors of the ith node to each other. Since the number of these links can range between 0 and $\frac{k_i(k_i-1)}{2}$, the average clustering coefficient can range between 0 and 1 and is therefore a rough measure of the connectedness of the graph. Now, for a BA type network with N nodes, the distribution of clustering coefficients goes approximately as $N^{-.75}$ whereas Ravasz et al. found that for the metabolic networks they studied the average clustering coefficient was independent of the network's size. To account for this, the authors proposed what they refer to as a 'hierarchical' network. The network is generated in the following manner. First, four completely connected nodes are taken as the starting point of a graph. These four nodes are then copied and and the three outer nodes of these copies are connected to the center of the original cluster (figure 4). This procedure is repeated a large number of times. First, for this graph $\gamma = 1 + \ln 4/\ln 3 \approx 2.26$ and has an average clustering coefficient $< C_i > \approx 0.6$ independent of the size of the network and therefore this model matches the data well. Of course [7] does not claim that this method of generating their hierarchical network has anything to do with the actual evolution of the network, just that the network constructed in this way roughly approximates the topology of the real metabolic networks. Of course the next highly non-trivial step is to understand what evolutionary processes led to this topology.

2.4 Spectral Properties

An alternative route to studying a network is to examine the spectral properties of its adjacency matrix. An adjacency matrix captures the complete connectivity of a network:

$$A_{ij} = \begin{cases} 1 \text{ if nodes } i \text{ and } j \text{ are connected} \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

The spectrum of eigenvalues $\rho(\lambda)$ of this matrix gives a great deal of information about it's topology. An amusing special case is that of a large random network. The distribution of eigenvalues turns out to be a semicircle!(see figure 5 and [1]) Eigenvalues give important dynamical information about processes that take place on a network. For example, the largest eigenvalue of an adjacency matrix determines how quickly an epidemic will spread [8]. Unfortunately, the total connectivity of many real-world networks is not known and even if it were calculating the spectrum of eigenvalues is often not practical. In light of this, [8] has studied *averageability* of extreme eigenvalues on scale-free networks. Specifically, they looked at ensembles of random scale-free networks and found that the extreme eigenvalues of the adjacency matrix have a small spread when averaged over the ensemble. Hence, the fact that the network is scale free puts tight constraints on what the largest eigenvalue is likely to be. The success of statistical mechanics is partly due to the fact that in the thermodynamic limit many distributions (energy for example) are sharply peaked about their mean. As the study

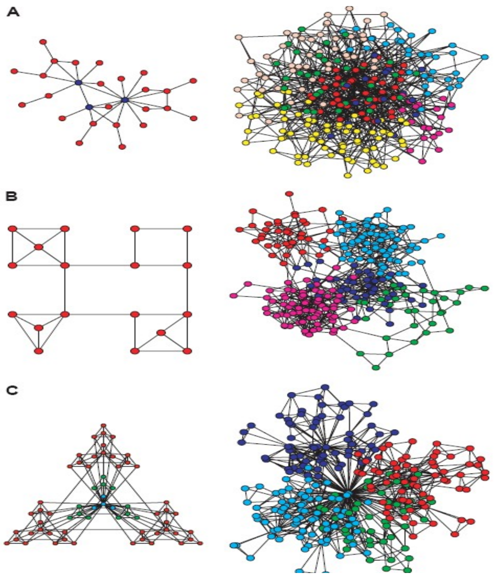

Figure 4: Examples of networks taken from [7]. On the left hand side are three schematic illustrations of a scale-free network, modular network, and hierarchical network respectively. On the right a much larger example of each of these networks is provided. For the scale-free network, highly connected nodes are blue. These hubs are instrumental in keeping the network together. The hierarchical network shows a scale-free topology but also embedded modularity. The different hierarchical levels are represented in increasing order from blue to green to red.

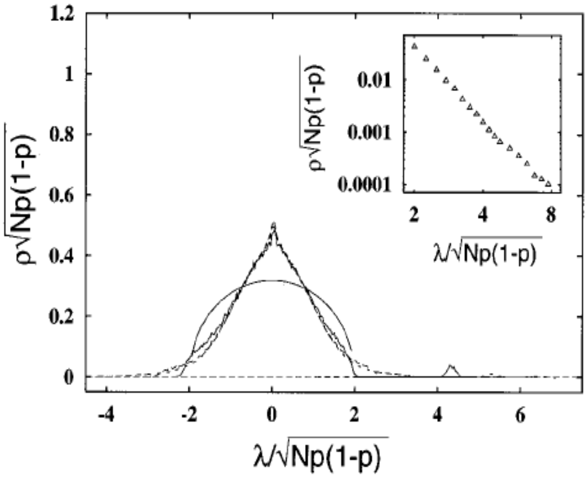

Figure 5: The spectral density of a random graph (semi-circular) and several BA generated networks of various sizes (tent-like). The inset shows the decay of the leading edge of the spectral density [1].

of networks broadens, the statistical analysis of networks requires sound theoretical grounding.

3 Network evolution

3.1 Preferential attachment

Of course the study of the statistical properties of graphs is a means to an end. A goal of network research is to gain insight into the local rules which dictate the large scale development of networks. Indeed, a great deal of network literature has focused on dynamics. We would not expect the simple BA model of linear preferential attachment to account for the wide variety of dynamical processes present in all scale free networks. First, the BA model gives $\gamma = 3$ [1] whereas we have seen that different networks have a wide range of exponents (see table 1). In addition, many real-world networks actually show deviations from simple power laws. For example most real-world networks have power law behavior with an exponential cutoff. One very natural explanation of cutoffs is that in many networks nodes have finite lifetime [1]. Another interesting effect is that in networks in which attachment preference $\Pi(k)$ can be directly measured it is often found that $\Pi(k)$ is a power law rather than a simple linear function of k (as in the BA algorithm).

3.2 Alternatives and Dissent

Preferential attachment is not the only possible explanation for the myriad of power laws seen in networks. [13] introduced a model in which locally interacting agents have partial and possibly incorrect information about the global network. Consider agents $i = 1, 2, ..., N$ which form the vertices of a network with E edges. Each agent has the following information about the network.

$$\begin{aligned} D_i(l) &= \text{agent } i\text{'s estimated shortest path to agent } l && (6) \\ P_i(l) &= \text{agent } i\text{'s nearest neighbor on the estimated shortest path to } l && (7) \end{aligned}$$

The dynamics are as follows.

(1) *Initial state*: The network starts with $N-1$ agents connected to a center agent. Additionally, there are $E-N+1$ randomly placed edges between the $N-1$ non-hub agents. All the information that each agent has is initially correct.

(2) *Loop*

(i) An agent i and one of its neighbors j are chosen at random.

(ii) A third agent l is chosen randomly and if $D_i(l) > D_j(l)$ the edge between i and j is rewired so that it connects i to $k = P_j(l)$.

(iii) Information that i has is lost and there is complete sharing of information between i and k.

Note that as rewirings take place, information that remote agents have becomes inaccurate. [13] also introduced a more complicated variant by introducing a parameter S. Now, the agent j receives a (tunable) fraction S of k's information. The global dynamics of the system are surprisingly sensitive to S. Above a critical value of S, a hub (the node with the largest degree) becomes 'frozen' and will never lose its central status. As S is lowered, at a critical value of S the location of the hub becomes dynamic and the degree distribution becomes scale free. For very low values of S, fluctuations become so large that no hubs develop and the resulting degree distribution is exponential. Not surprisingly, a more efficient exchange of information has a stabilizing effect on the system. Although [14] primarily focuses on the effects of preferential attachment on the evolution of Wikipedia, the idea of local information exchange was invoked as a possible secondary effect.

Doyle and Carlson have taken issue with the application of self-organized criticality to certain designed systems, most specifically the internet. The idea behind *highly optimized tolerance* (HOT) is that self-organization is often not present in a carefully engineered system. For example, it is argued in [9] that large fluctuations are not a dominant effect in internet traffic patterns because sophisticated protocols have been specifically designed for flexibility and robustness and act to suppress such fluctuations. Although a full exploration of the HOT ideas would take us too far afield, suffice it to say that the objections of Doyle and Carlson are inapplicable to the types of networks we've studied here. For example, the topology of web page connections is certainly not optimized in the HOT sense.

3.3 Directed Graphs

For the most part, we have been discussing undirected graphs. However, many networks like the internet are really directed graphs: web page links have a clear direction associated with them. As we will see, ignoring the directional nature of a network may actually lead to spurious conclusions! As pointed out in [11], this effect is an example of *Simpson's paradox*: correlations can change if different types of data are pooled. Consider the following example [15]. In 1995, Derek Jeter had a batting average of 0.250 (12/48) while David Justice ended the year with 0.253 (104/411). Now, in 1996 Jeter was 0.314 (183/582) while David Justice had an impressive 0.321 (45/140). Note that the combined batting average for the two years is Jeter 0.310 (195/630) while Justice is at 0.270 (149/551)! The trend has reversed! Of course, there is no real paradox involved here. But the point is clear: pooling data can lead to spurious conclusions.

In [12], Newman looked at the internet as a simple undirected network and found no correlation between the degrees of nearest neighbors. This would rule out preferential attachment as a mechanism since preferential attachment leads to a positive correlation in the degrees of nearest neighbors. However, [11] took the directionality of the internet into account and they found that the node degree of a website correlated positively with the degree of both upstream and downstream neighbors (figure 6). This study provides an interesting cautionary note: models which ignore the directionality of graphs should be regarded with suspicion.

4 Future Directions

As we have seen, two networks may have very different qualitative topologies even though both are scale-free. Unlike critical exponents (in other areas of statistical mechanics) which are often dictated by fundamental aspects of the problem (it's dimensionality, for example), the exponents (γ for example) are in general tunable and very similar networks can give rise to a range of γ's [1]. For a power law preferential

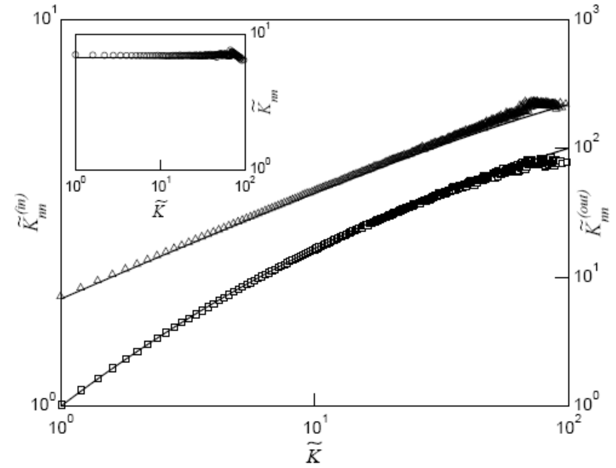

Figure 6: $\widetilde{K}_{nn}^{in}$ and $\widetilde{K}_{nn}^{out}$ measure the average degrees of upstream and downstream neighbors respectively. The large graph is a histogram of $\widetilde{K}_{nn}^{in}$ (squares) and $\widetilde{K}_{nn}^{out}$ (triangles) plotted against $\widetilde{K}$, the average degree (incoming and outgoing links combined) of the various nodes. Both the degree of upstream and downstream neighbors show a positive correlation with the degree of the given node. The inset shows that if the data is pooled the correlation disappears.

attachment ($\Pi \approx k^{\alpha}$), changing α would lead to a change in γ. Also, very different models can give power laws with exactly the same exponent. Of course this is a common problem in condensed matter problems. There are often many putative models which can lead to similar behavior.

But the situation is not hopeless. As we have seen, there are many studies which have gone a long way in helping explain the structure and, to a lesser extent, the origin of networks. What are the ingredients common to these successful studies? The first step in studying a proposed network is to get topological data. Whether the raw data has topology built into it (as in [**?**]) or must be inferred (as in [5]) it's crucial that the various metrics are not sensitive to possible errors in the data or details of how the underlying network is extracted. For example, if γ in [5] were sensitive to the parameter μ, which value to take would become a nontrivial issue. Additionally, many different underlying dynamics can give rise to power laws, so it's crucial to use several different metrics (such as the clustering coefficient) to give a more complete picture of the topology. Any proposed model that is built to explain the data must account for the various metrics. Better yet, if additional information about the nodes is available (for example their relative age) it is crucial that this information be used to corroborate the model. But a model which simply *describes* the connectivity can not be the ultimate goal. Of course, a description of a disease network may go along way in helping to combat the disease, but as far as our intellectual curiousity is concerned studies must go farther. That is, the model must be used to gain insight into how the network developed and how interactions hold it together.

The emerging study of networks shows great promise. As pointed out in [7], it would be very exciting to know what evolutionary processes led to the hierarchical metabolic structure we see today. One of the challenges in studying the origins of life is that as life is constantly evolving information is ostensibly lost. Perhaps the study of modern biological networks will provide a means to look back and learn something about our earliest ancestors.

References

[1] Albert-Laszlo and Barabasi. Reviews of Modern Physics, 74, 47-97 (2002).

[2] Barabasi, Albert-Laszlo. "Scale-Free Networks" *Scientific American*, 288:60-69, (2003).

[3] Cohen and Havlin. "Scale-Free Networks are ultrasmall". PRL, 90, 5, 058701. (2003).

[4] May and Lloyd. "Infection dynamics on a scale-free network". Phys. Rev. E, 64, 066112. (2001).

[5] Small, Walker and Tse. "Scale-Free Distribution of Avian Influenza Outbreaks". PRL 99, 188702. (2007).

[6] Jeong, Tombor, Albert, Oltvai, Barabasi. "The large-scale organization of metabolic networks". *Nature*, 407. October 2000.

[7] Ravasz, Somera, Mongru, Oltvai, Barabasi. "Hierarchical Organization of Modularity in Metabolic Networks". *Science*, 297. August 2002.

[8] Kim and Motter. "Ensemble Averageability in Network Spectra". PRL, 98, 248701. (2007).

[9] Carlson and Doyle. "Highly Optimized Tolerance: Robustness and Design in Complex Systems". PRL, 84, 2529. (2000).

[10] Stryer. *Biochemistry*. 3rd Ed. W.H. Freeman and Company. 1975.

[11] Capocci and Colaiori. "Mixing properties of growing networks and Simpson's paradox". arXiv:cond-mat/0506509v2. (2005)

[12] Newman. "Assortive mixing in networks". arXiv:cond-mat/0205405v1. (2002).

[13] Rosvall and Sneppen. "Modeling Dynamics of Information Networks". PRL 91, 178701. Oct. (2003).

[14] Capocci et. al. "Preferential attachment in the growth of social networks: the case of Wikipedia". arXiv:physics/0602026v2. (2006).

[15] The example of batting averages is taken from the English Wikipedia article on "Simpson's Paradox". Of course, in this example it's not at all clear that it's better to look at the batting averages from individual years. However, the point is that if we expect that pointing to or pointing from a web page have different mechanisms, it is imperative to seperate this data.