W.5.7 Sharabati dissertation and [SAI2010]

W.5.7.1 Plagiarism and scholarship concerns

[**SHA2008**] Walid Sharabati, *Multi-Mode and Evolutionary Networks*, 10/31/08. PhD Dissertation, GMU. (co-supervised by Said and Wegman) digilib.gmu.edu:8080/handle/1920/3384?mode=full

This 230-page dissertation shows much work. Some comes from [SHA2006], which is fine. It also uses several plagiarized sources,[75] and has other basic scholarship problems.

His Committee might have guided him more. One might worry that in-depth literature review got lost in an attempt to cover too many topics.

pp.1-3

The text here is described in **W.2.3**, derived from other antecedents, [WIK2006a, WAS1994, DEN2055], i.e., Wikipedia, Wasserman and Faust (1994), and de Nooy, Mrvar, Batagelj (2005), used in various publications.

p.8

Several paragraphs are near-verbatim extracts from [HAN2005], **W.5.7.2**.

p.9

> "I conjecture based on two papers published recently [55, 56] that certain styles of co-authorship lead to the possibility of group-think, reduced creativity, and the possibility of less rigorous reviewing processes."

This again repeats Meme-b1, with zero supporting evidence.

Did no one on Committee question this unsupported claim?

The two references are:

> "[55] Y. Said, E. Wegman, W. Sharabati, and J. Rigsby, Implications of co-author networks on peer review, (2007).
> *That is in Classification and Data Analysis, Macerata, Italy: EUMEdizoni Università di Macerata, 245-248, 2007. I have found no online copy.*
>
> [56] ____, Social networks of author-coauthor relationships, Computational Statistics and Data Analysis 52 (2007), 2177{2184, DOI: 10.1016/j.csda.2007.07.021."
> This called [SAI2008} here to match the final publication year.

p.9-10

> "Of all the work that has been done on social networks, very few scientists had considered coauthorship networks. The main goal of analyzing coauthorship networks is to be able to answer the question of who-wrote-with-whom" and with what frequency.

This appears in [SHA2006] in an even stronger form, discussed in **W.5.3** in more detail, Theme-M4 *Every PhD likes to think their work opens new areas. Did his Committee really believe this claim?*

pp.124-125

Several paragraphs are near-verbatim extracts from [HAN2005], apparently via a few more edits to the text on p.8, see **W.5.7.2**.

pp.128-129

Several paragraphs are near-verbatim extracts from a famous paper, [BAR1999] and a Wikipedia page that points there [WIK2007], **W.5.7.3**. These are especially noteworthy as they also appear in the later [SAI2010], which also cites 3 government contracts, **W.5.7.4**.

References used only in this section (for now) are:

[BAR1999] Albert-László Barabási and Réka Albert, "Emergence of Scaling in Random Networks," *Science* 15 October 1999: Vol. 286 no. 5439 pp. 509-512 DOI: 10.1126/science.286.5439.509.[76]

[HAN2005] R. Hanneman and M. Riddle, Introduction to social network methods, Online textbook: Riverside, CA 2005. http://www.faculty.ucr.edu/_hanneman/nettext

[**SAI2010**] Yasmin H. Said, Edward J. Wegman, andWalid K. Sharabati, "Author–Coauthor Social Network and Emerging Scientific Subfields," F. Palumbo et al. (eds.), *Data Analysis and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-642-03739-9_30, ©Springer-Verlag Berlin Heidelberg 2010, pp.257-268.

[WIK2007] Wikipedia page, edit 04/08/07 by AnAj.[77]

[75] www.desmogblog.com/mashey-report-reveals-wegman-manipulations
See p.7, note (3) where Wegman explains giving Reeves' work to Sharabati, "as background material along with a number of other references. Walid included it as background material in his dissertation with only minor amendments**."** Possibly, some of the "other references" are those shown later.

[76] http://www.sciencemag.org/content/286/5439/509.full?sid=65cf560b-7c0a-4a4d-93cb-ca8cb80a0d85 [subs. Req]

[77] en.wikipedia.org/w/index.php?title=Barab%C3%A1si%E2%80%93Albert_model&oldid=121157236

I am also surprised no one noticed the weakness of reference and citation:
pp.209-213 Bibliography
Of 67 references only 26 seemed to be cited, with page numbers shown.
GMU or Wegman-related sources are Bold, leaving 20 others.

[2]	129	Barabasi, Albert
[5]	10	Borner
[6]	129	Borner
[8]	15	Carley
[11]	130	**Cioffi-Revilla (**GMU**)**
[12]	31,129	CIS
[13]	129	DBLP
[14]	5	De Nooy, Mrvar, Batelgj
[21]	31,146	FARS
[24]	82	Gile and Handcock
[28]	110	Hanneman and Riddle
[30]	26	Seock-Ho Kim
[34]	34	Krackhardt and Carley
[41]	22,23	**Marchette** and **Priebe**
[44]	24	Mukha
[48]	129	PubMed
[49]	24	Robertson
[50]	24	Robins, Pattison, Kalish, Lusher
[51]	129	Roth
[53]	11,13,21	Said
[55]	9,18,29	Said**,** Wegman**, Sharabati,** Rigsby
[56]	9,18	Said**,** Wegman**, Sharabati,** Rigsby
[57]	24	Simpson
[60]	24	van Duijn, Gile, Handcock
[61]	4,6,6,7,135	Wasserman and Faust
[62]	11,12	Wegman and Said
no ref	27	Mielke (1978) (cited, but no reference)
no ref	27	Faust and Romney (1985)
no ref	31,75	**Martinez** (2002)

The WR cites ~ 50% of its references. [SHA2008] cites ~40%, although a few may have been missed. For the range of topics covered, the number of references seems low, as the co-authorship literature alone could easily include that many. Many references lacked sources or page numbers.

Given all this, it seems the Bibliography was not examined with much care by the Committee (Wegman, Said, Robert Axtell, Igor Griva, Tim D. Sauer, Maxim Tsvetovat.)
This dissertation got "best departmental dissertation of year" award.

Now, we return to the details of the recently-identified plagiarism flows.
Wegman's February 2010 C.V. listed:

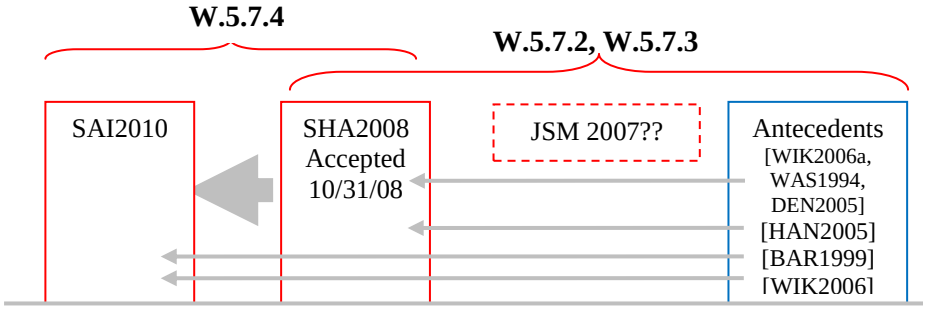
> 179. "Style of author-coauthor social networks," with Yasmin H. Said, Walid K. Sharabati, John T. Rigsby, *Computational Statistics and Data Analysis*, 52, 2177-2184, 2008; doi:10.1016/j.csda.2007.07.021, 2007

That was [SAI2008], but wrong title.

> 183. "A model of preferential attachments for emerging scientific subfields," with Walid K. Sharabati and Yasmin H. Said, *Proceedings of the Joint Statistical Meetings*, 2048-2055, 2007.

JSM only mentions Sharabati for Sharabati, Said, Wegman, "Style of Author-Coauthorship Social Networks: Statisticians of Prominent U.S. Universities,"[78] so some may have appeared then.

Following is the apparent flow of SNA-related material



[SAI2010, p.267] has the same Acknowledgements as [SAI2008]:

> **"Acknowledgements** The work of Dr. Said is supported in part by Grant Number F32AA015876 from the National Institute on Alcohol Abuse and Alcoholism. The work of Dr. Wegman is supported in part by the Army Research Office under contract W911NF-04-1-0447. Both were also supported in part by the Army Research Laboratory under contract W911NF-07-1-0059.

The only affiliation given for the 3 authors was:

> "Y. H. Said, Isaac Newton Institute for Mathematical Sciences, Cambridge University, Cambridge, CB3 0EH UK e-mail: EMAIL and
> Department of Computational and Data Sciences, George Mason University MS 6A2, Fairfax, VA 22030, USA"

[78] www.amstat.org/meetings/jsm/2007/onlineprogram/index.cfm?fuseaction=abstract_details&abstractid=308858

W.5.7.2 [SHA2008, pp.124-125 ← p.8] ← [HAN2005]

[SHA2008, p.124-125] [79]

The purpose of equivalence analysis is to identify "classes" or clusters based on similarity.
I implicitly assume that distances among actors reflect as a two dimensional; although, it is possible that the data are multi-dimensional.

MDS is used (metric for data that are inherently valued) to cluster actors based on distance.

MDS represents the patterns of similarity or dissimilarity in the tie profiles among actors (when applied to adjacency or distances) as a "map" in multi-dimensional space. The map lets us see how "close" actors are, whether they "cluster" in multi-dimensional space and how much variation there is along each dimension.
The goal of MDS is to minimize stress - distance between nodes.
"Stress" is a measure of badness of fit; $0 \leq$ stress ≤ 1.

The range of solutions with more dimensions is sought, so that the analyst can assess the extent to which the distances are uni-dimensional. The meaning of the dimensions can sometimes be assessed by comparing agents that are at the extreme poles of each dimension.

The flow seems fairly clear, from Hanneman and Riddle, to [SHA2008, p.8] and then edited further for [SHA2008, p.124-125].

[SHA2008, p.8]

The purpose of equivalence analysis is to identify and visualize "classes" or clusters. In cluster analysis, it is implicitly assumed that the similarity or distance among actors reflects as single underlying dimension. It is possible, however, that there are multiple "aspects", "attributes" or "dimensions" underlying the observed similarities of cases.
Components analysis could be applied to correlations among actors. Alternatively, MDS (Multi-Dimensional Scaling) could be used (metric for data that are inherently valued) to cluster the actors.

MDS represents the patterns of similarity or dissimilarity in the tie profiles among the actors (when applied to adjacency or distances) as a "map" in multi-dimensional space. This map lets us see how "close" actors are, whether they "cluster" in multi-dimensional space, and how much variation there is along each dimension.
The aim of MDS is to minimize stress - distance between vertices.
"Stress" is a measure of badness of fit; $0 \leq$ stress ≤ 1. In MDS, we look at a range of solutions with more dimensions, so we can assess the extent to which the distances are uni-dimensional. The "meaning" of the dimensions can sometimes be assessed by comparing agents that are at the extreme poles of each dimension.

The cases above seems to be a missing change to actors or agents.

[HAN2005] specifically following section:
http://www.faculty.ucr.edu/~hanneman/nettext/C13_%20Structural_Equivalence.html
This is [SHA2008] ref. [28] ,cited only on p.110, nowhere near either usage at left.

Usually our goal in equivalence analysis is to identify and visualize "classes" or clusters ~~of cases~~. In ~~using~~ cluster analysis, we are implicitly assuming that the similarity or distance among cases reflects as single underlying dimension. It is possible, however, that there are multiple "aspects" or "dimensions" underlying the observed similarities of cases. ~~Factor or~~ components analysis could be applied to correlations ~~or covariances~~ among cases. Alternatively, multi-dimensional scaling could be used (~~non-~~metric for data that are inherently ~~nominal or ordinal; metric for~~ valued).

MDS represents the patterns of similarity or dissimilarity in the tie profiles among the actors (when applied to adjacency or distances) as a "map" in multi-dimensional space. This map lets us see how "close" actors are, whether they "cluster" in multi-dimensional space, and how much variation there is along each dimension. …

"Stress" is a measure of badness of fit. In ~~using~~ MDS, it is a good idea to look at a range of solutions with more dimensions, so you can assess the extent to which the distances are uni-dimensional. …The "meaning" of the dimensions can sometimes be assessed by comparing cases that are at the extreme poles of each dimension.

I have not done a serious search for more re-uses of [HAN2005].

[79] http://deepclimate.org/2011/05/16/retraction-of-said-wegman-et-al-2008-part-2/#comment-9091
May 30, 2011, thanks to *andrewt*.

W.5.7.3 [SHA2008, pp.128-129] ← [BAR199, WIK2007]

*May 29, 2011, andrewt showed that these two unacknowledged sources appeared in the article [SAI2010] discussed later in **W.5.7.4**.*

Further investigation found that these had been incorporated in to [SHA2008] first, and then [SAI2010] was derived entirely from the dissertation, with minor edits. The black-outlined paragraphs appear near-verbatim in the latter. [80]

[SHA2008, p.128]

5.3 A Model of Preferential Attachment for Emerging Scientific Subfields

In this section, I focus on demonstrating scale-free author-coauthor social networks. A common property of many large networks is that the vertex connectivities follow a scale-free power-law distribution. This feature was found to be a consequence of two generic mechanisms: (i) networks expand continuously by the addition of new vertices (growth), and (ii) new vertices attach preferentially to sites that are already well connected (preferential attachment). A model based on these two ingredients reproduces the observed stationary scale-free distributions, which indicates that the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems.

[BAR1999, p.509]

'A common property of many large networks is that the vertex connectivities follow a scale-free power-law distribution. This feature was found to be a consequence of two generic mechanisms: (i) networks expand continuously by the addition of new vertices, and (ii) new vertices attach preferentially to sites that are already well connected. A model based on these two ingredients reproduces the observed stationary scale-free distributions, which indicates that the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems.'

Growth means that the number of vertices (actors) increases with time.
Preferential attachment means that the more connected a vertex is, the more likely it is to acquire new edges.'

[WIK2007]Barabasi-Albert Model

'Growth means that the number of nodes in the network increases over time.
Preferential attachment means that the more connected a node is, the more likely it is to receive new links.

Intuitively, preferential attachment can be understood if we think in terms of social networks connecting people. Here an edge from actor A to actor B means that actor A

Intuitively, ~~the~~ preferential attachment can be understood if we think in terms of social networks connecting people. Here a link from A to B means that person A

[SHA2008, p.129]

"knows" or "is acquainted with" actor B. Vertices with many edges represent well-known people with lots of relations. When a new actor enters the community, he or she is more likely to become acquainted with one of those more visible actors rather than with a relative unknown.

"knows" or "is acquainted with" person B. Heavily linked nodes represent well-known people with lots of relations. When a newcomer enters the community, s/he is more like to become acquainted with one of those more visible people rather than with a relative unknown.'

Models that satisfy these two principles are known as Barabsi-Albert models [2].

Barabasi is spelled correctly in the Bibliography. Barabasi-Albert-derived text is separated from the citation above by intervening Wikipedia text.

[WIK2007] cites [BAR1999], so it seems plausible to have started with Wikipedia and found the reference, and then cited the latter, but not the former.

[80] http://deepclimate.org/2011/05/16/retraction-of-said-wegman-et-al-2008-part-2/#comment-9063

W.5.4 [SAI2010] ← [SHA2008]

Sharabati used several paragraphs from [HAN2005, BAR1999, WIK2007] in his dissertation [SHA2008], adding to several pages given to him by Wegman, who got them from Denise Reese.[81]

Then, a year later, it appears that [SHA2008, §5.3, pp.128-144] was turned into a conference paper, and later published in the proceedings.

- Subsections were reordered somewhat.
- Minor edits made some improvements, some marginal.
- "I" was changed to "we" everywhere.
- Citations were fixed to match the journal style.
- Some references were made more precise.
- **[SHA2008] was not referenced or cited.**
- **The authorship was Said, Wegman, Sharabati.**

Sharabati is an author, so this transformation might not be considered plagiarism.[82] PhD students often (and reasonably) extract parts of their dissertation as articles, and sometimes supervisors get added as coauthors, also often just fine.
I have no idea how common it is for 2 co-supervisors' names to appear ahead of the student for material almost entirely taken, near-verbatim, from his dissertation, with no new work beyond minor editing.

Some people would not consider this very good supervision practice. Of course, I have no idea who did the actual work, perhaps Sharabati did and was grateful for another publication.

Once again, the same 3 Federal agencies were acknowledged for funding for Said or Wegman. Presumably, they included this paper their reports to the agencies, assuming any have yet been filed on this.

On following pages, the side-by-sides show the entire final article text at left, with antecedent text from the dissertation at right, minus illustrations, of which the former are subset of the latter.

The portions outlined in black seem to be derived from [BAR1999, WIK2007], as per W.5.7.3.

[81] www.desmogblog.com/mashey-report-reveals-wegman-manipulations, p.7.
[82] Of course it copied the [BAR1999, WIK2007] problems.

[SAI2010, p.257]
'**Abstract** In this paper, we suggest a model of preferential attachment in coauthorship social networks.

The process of one actor attaching to another actor (author) and strengthening the tie over time is a stochastic random process based on the distributions of tie-strength and clique size among actors. We will use empirical data to obtain the distributions.
The proposed model will be utilized to predict emerging scientific subfields by observing the evolution of the coauthorship network over time. Further, we will examine the distribution of tie-strength of some prominent scholars to investigate the style of coauthorship.
Finally, we present an example of a simulated coauthorship network generated randomly to compare with a real-world network.

[SHA2008, p.21]
'To conclude, I present a mathematical model of preferential attachment in coauthorship socio-networks.'

[SHA2008, p.129]
'The process of one actor attaching to another actor (author) and strengthening the tie over time is a stochastic random process based on the distributions of tie-strength and clique size among actors, which are obtained from empirical data.
I then utilize the model to predict emerging scientific subfields of the evolutionary coauthorship network.
Followed by a discussion on style of coauthorship among prominent scholars that is using the distribution of tie-strength…."
I first present the model that is based on the theory presented in Chapter 4 and then compare a randomly generated network with a real network."'

1 Introduction

[SHA2008, p.128]
5.3 A Model of Preferential Attachment for Emerging Scientific Subfields

From [BAR1999]

"In this paper, we focus on demonstrating scale-free author–coauthor social networks. A common property of many large networks is that the vertex connectivities follow a scale-free power-law distribution. This feature was found to be a consequence of two generic mechanisms: (1) networks expand continuously by the addition of new vertices (growth), and (2) new vertices attach preferentially to sites that are already well connected (preferential attachment). A model based on these two ingredients reproduces the observed stationary scale-free distributions, which indicates that the development of large networks is governed by robust selforganizing phenomena that go beyond the particulars of the individual systems.

In this section, I focus on demonstrating scale-free author-coauthor social networks. A common property of many large networks is that the vertex connectivities follow a scale-free power-law distribution. This feature was found to be a consequence of two generic mechanisms: (i) networks expand continuously by the addition of new vertices (growth), and (ii) new vertices attach preferentially to sites that are already well connected (preferential attachment). A model based on these two ingredients reproduces the observed stationary scale-free distributions, which indicates that the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems.

From [WIK2007]

Growth means that the number of vertices (actors) increases with time. Preferential attachment means that the more connected a vertex is, the more likely it is to acquire new edges.'
'Intuitively, preferential attachment can be understood if we think in terms of social networks connecting people. Here an edge from actor A to actor B means that actor A
[SAI2010, p.258]
"knows" or "is acquainted with" actor B. Vertices with many edges represent well-known people with lots of relations. When a new actor enters the community, he or she is more likely to become acquainted with one of those more visible actors rather than with a relative unknown. Models that satisfy these two principles are known as Barabasi–Albert models (Barabasi and Albert 1999). In this paper, we seek to demonstrate that author–coauthor networks in the statistical literature satisfy these two criteria.

Growth means that the number of vertices (actors) increases with time. Preferential attachment means that the more connected a vertex is, the more likely it is to acquire new edges.'
'Intuitively, preferential attachment can be understood if we think in terms of social networks connecting people. Here an edge from actor A to actor B means that actor A
[SHA2008, p.129]
"knows" or "is acquainted with" actor B. Vertices with many edges represent well-known people with lots of relations. When a new actor enters the community, he or she is more likely to become acquainted with one of those more visible actors rather than with a relative unknown. Models that satisfy these two principles are known as Barabasi-Albert models [2]. In this section, I seek to demonstrate that author-coauthor networks in the statistical literature satisfy these two criteria.

[SAI2010, p.258, cont]
There has been work on author-coauthor networks and the emergence of global brain in Borner et al. (2005), preferential attachment in Roth (2005), and implications for peer review in Said et al. (2008). Coauthorship relationships can be treated as a two-mode networks in which there are two types of nodes; the author nodes and paper nodes, and one relationship type; "person A authored/coauthored paper P"

This two-mode social network is expressed in the PCANS model (Krackhardt and Carley 1998; Carley 2002). The PCANS model is represented in the table below:"

There is a one-to-one correspondence between graphs and matrices; a graph can be fully represented using a matrix. Moreover, matrix algebra is well-defined. Therefore, we will use matrix operations to obtain new socio-matrices having new properties. Consider the two-mode "author-by-paper" binary social network AP, then
AP x AP^T = AP x PA = AA;
is the one-mode network of authors related through papers. Similarly,
AP^T x AP = PA x AP = PP;
is the one-mode network of papers related through authors.

The author-by-author socio-matrix AA is one of interest because it exhibits relationships among authors, in other words, the author-by-author matrix tells "who-wrote-with-whom".'

Data on statisticians and statistics subfields were collected from the online Current Index to Statistics (CIS) database
The procedure used to harvest data involved two stages. First, we queried the database using names of well-established statisticians affiliated with prominent US universities. These data were used to build a

[SAI2010, p.259]
social network of coauthors and to derive the distribution of tie-strength "frequency of coauthorship" among coauthors. A different dataset was used to derive the distribution of clique size. In the second stage, we used the

biopharmaceutical subfield as a keyword to query the database, the dataset was used to discover the emergence of that scientific subfield by exploring the evolution of the coauthorship social network over time as a time series.

[SHA2008, p.129, cont]
There has been work on author-coauthor networks and the emergence of global brain in [6], preferential attachment in [51] and implications for peer review in [55]. Coauthorship relationships can be treated as a 2-mode networks in which there are two types of nodes; the authors nodes and the papers nodes, and one relationship type; "person A authored coauthored paper P".

This two-mode ~~relational~~ socio-network can be concluded from the PCANS model [34], [8]. Table 1.1 portrays the PCANS model.

I can perform matrix operations such as the product of matrices to obtain interesting results given that the two-mode matrix is binary. Let the two-mode "author-by-paper" binary social matrix AP be given, then
AP x AP^T = AP x PA = AA;
is the one-mode network of authors related through papers. Similarly,
AP^T x AP = PA x AP = PP;
is the one-mode network of papers related through authors.

The author-by-author social matrix AA is one of interest, it reveals relationships among authors, in other words, the author-by-author matrix resembles the "who-wrote-with-whom" relationship.'

Data on statisticians and statistics subfields were collected from the online Current Index to Statistics (CIS) database ~~[12]~~.…
The procedure used to harvest data involved two stages using names of well-established statisticians affiliated with prominent US universities. These data were used to build a

social network of coauthors and to derive the distribution of tie-strength "frequency of coauthorship" among coauthors. A different dataset was used to derive the distribution of clique size. In the second stage, I used the

[SHA2008, p.130]
biopharmaceutical as keywords to query the database, the dataset was used to discover the emergence of scientific subfields by exploring the evolution of the coauthorship socio-networks over time as a time series.

[SAI2010, p.259, cont]
2 Distribution of Tie Strength
In weighted coauthorship social networks, strength of a tie indicates the frequency of coauthored papers between two actors; in other words, it is a measure of how close two actors are and how much they trust each other. Therefore, studying tie-strength is a subject of interest in coauthorship social networks. We developed a MATLAB program to build the one-mode proximity matrix of the data collected from the CIS database on contributing scientists in the field of statistics. This adjacency weighted matrix was later manipulated to construct the distribution of tie-strength. The statisticians dataset contained 1,767 published papers that had 874 unique author(s)/coauthor(s), the one-mode network of coauthors is shown in Fig. 1.

The distribution of tie-strength is shown in Fig. 2.

'Figure 2 suggests a power law distribution (Cioffi-Revilla 2005).

Because the density curve is close to linear in log-log space the distribution is power

law, the next step would be computing the exponent α of the power law. This can be done either by finding the slope of the least-squares regression line in log-log space or by
[SAI2010, p.260]
using the following aggregation method for calculating the exponent α .

(identical equations and text, until
Therefore, we can observe that the distribution of tie-strength is power law with exponent value of 2.17.

Looking into the low-level processes that produced the many-some-few power law pattern, we conjecture that this behavior can be generated in view of the following reasons. First of all, there

are higher chances to find two coauthors who simply published together few times.

Many of these statisticians are professors who may have a number of graduates working with on a project or paper at a given time period. Upon graduation, many of these students prefer a career in the industry, therefore, they lose contact with their professors leaving behind one or two published papers with that professor. On the other hand, some scientists find themselves in the research area, as a result, the likelihood that two already coauthored individuals publish again rises. If you coauthored a good quality paper with someone and you liked him/her, chances you are going to publish with him/her again if there is mutual agreement increase. And finally, there are those authors who favor only very few coauthors; a colleague or a fellow student who maintains good contacts and relations with that author, to publish with the most.

[SHA2008, p.130, cont]
5.3.1 Distribution of Tie Strength
In weighted coauthorship socio-networks, strength of a tie indicates the frequency of coauthored papers between two actors; in other words, it is a measure of how close two actors are and how much they trust each other. Therefore, studying tie-strength is a subject of interest in coauthorship social networks. We developed a MATLAB program to build the 1-mode proximity matrix of the data collected from the CIS database on contributing scientists in the field of statistics. This adjacency weighted matrix was later manipulated to construct the distribution of tie-strength. The statisticians dataset contained 1767 published papers that had 874 unique author(s)/coauthor(s), the 1-mode network of coauthors is shown in Figure 5.32, …

The distribution of tie-strength is shown in Figure 5.35(a)

Figure 5.35(a) suggests a power law distribution [11]. ~~To investigate this, I first plotted the distribution in log-log scale, this is shown in Figure 5.35(b).~~
Because the density curve is close to linear in log-log space, ~~it is reasonable to conjecture that~~ the distribution is power
[SHA2008, p.134]
law. The next step would be computing the exponent α of the power law. This can be done by either finding the slope of the least-squares regression line in log-log space or by

using the following aggregation method for calculating the exponent α.

(identical equations and text, until
Therefore, I can observe that the distribution of tie-strength is power law with exponent value of approximately 2.1716.

Looking into the low-level processes that produced the many-some-few power law pattern, I conjecture that this behavior is generated in view of the following reasons. Firstly, there
[SHA2008, p.135]
are higher chances to find two coauthors who simply published together few times ~~or perhaps once.~~
Many of these statistician are professors who may have a number of graduates working on projects or papers at a given time period. Upon graduation, many of these students prefer a career in the industry, therefore, they lose contact with their professors leaving behind one or two published papers with that professor. On the other hand, some scientists find themselves in the research area, as a result, the likelihood that two already coauthored individuals publish again rises. If you coauthored a good quality paper with someone and you liked him/her, chances are you are going to publish with him/her again if there is mutual agreement increase. And finally, there are those authors who favor only very few coauthors; a colleague or a fellow student who maintains good contacts and relations with that author, to publish with the most.

[SAI2010, p.260, cont]
We further investigated the distribution of tie-strength of individual authors. Figure 3 shows a typical distribution of tie-strengths. We investigated three additional authors.
Surprisingly, the distribution is again power-law with exponent α

[SAI2010, p.261]
ranging 1.5–1.85. Because α <2 both the mean and the variance of the distribution of the power-law are not defined and hence the power-law is said to be not well-behaved. For the mean and variance of a power-law to be well-behaved α ¸ has to be greater than 3, if 2 < α < 3 only the mean is finite. We also note that the distribution of tie-strength is a self-similar power-law distribution for coauthorship social networks.

Distribution of Clique Size
An important factor in preferential attachment is the clique size; the number of people coauthored a single paper. Note that a paper with sole author or two coauthors is technically not considered a clique. A clique in a graph must have at least three fully connected nodes "complete graph/subgraph" Wasserman and Faust (1994).
The statisticians dataset was used to construct
the distribution of clique size to obtain a better understanding of how coauthors interact.

Figure 4 shows the distribution of

[SAI2010, p.262]
clique size. The distribution of clique size is approximately lognormal with mean μ = 1:954 and standard deviation σ = 1:6.

4 Random Graph Model for Preferential Attachment
The model is based on stochastic "random" processes, in which nodes are generated randomly at each time step. At each time step, a new paper gets published and one of three things could happen:
1. New actor(s) try to attach to existing actors.
2. Already existing non-attached actor(s) attempt to make an attachment(s).
3. Already attached actor(s) strengthen their ties.

[SAI2010, p.262]
And each node has the attributes:
1. Name
2. Age
3. Weight
4. Preference
5. Status
6. Field
7. Active flag

[SHA2008, p.135, cont]
I further investigated the distribution of tie-strength of individual authors. Figure 5.36 shows the distribution of tie-strength of four different authors
Surprisingly, the distribution is again power-law with exponent α

ranging 1.5–1.85. Because α <2 both the mean and the variance of the distribution of the power-law are not defined and hence the power-law is said to be not well-behaved. For the mean and variance of a power-law to be well-behaved α ¸ has to be greater than 3, if 2 < α < 3 only the mean is finite. The distribution of tie-strength is a self-similar power-law distribution for coauthorship social networks.

5.3.2 Distribution of Clique Size
An important factor in preferential attachment is the clique size; the number of people coauthored a single paper. Note that a paper with sole author or two coauthors is technically not considered a clique. A clique in a graph must have at least three fully connected nodes \complete graph" [61].
I used the dataset of prominent statisticians to construct
the distribution of clique size to obtain a better understanding of how coauthors interact

[SHA2008, p.137]
Figure 5.37 shows the distribution of

clique size. The distribution of clique size is approximately lognormal with mean μ = 1:954 and standard deviation σ = 1:6.

[SHA2008, p.138]
5.3.4 Random Graph Model
The model is based on stochastic "random" processes, in which vertices are generated randomly at each time step. At each time step, a new paper gets published and one of three things could happen.
1. New actor(s) try to attach to existing actor(s).
2. Already existing non-attached actor(s) attempt to make an attachment(s).
3. Already attached actor(s) strengthen their ties.

[SHA2008, p.138]
And each vertex has the attributes:
name - age – weight - preference - status field -
active flag.

[SAI2010, p.262, cont]
These attributes uniquely identify actors, some of which change rapidly/slowly over time while other attributes remain the same over time. For example, the attributes “name” and “field” do not change. The evolution of “weight” and “status” attributes can be viewed as a time series because they change faster than any other attributes. “Age” changes linearly over time. Meanwhile, the “active” flag operates as a switch initially set to “on” but later could change to “off”, once it is changed to “off” it remains in that state forever. Certain actors might change the attribute “preference”.

The model was implemented in MATLAB and consists of approximately 350 lines of code, it exploits the distributions of tie-strength and clique-size to build the coauthorship network. Figure 5 is a two-mode author-by-paper simulated network.

[SAI2010, p.263]
Note that a new publication surfaces at each time step. Figure 6 shows the one-mode coauthorship network corresponding to the matrix in Fig. 5.

Figure 7 shows a simulated coauthorship social network, the program ran for 100 iterations. The simulated network is similar to the network obtained from empirical data, see Sect. 5.

5 The Emergence of Scientific Subfields
Here we explore
The social network of biopharmaceutical statisticians over time to inspect the emergence of this subfield. The data include papers published between the years 1977 and 2003. There are 157 published papers with 260 unique author(s)/ coauthor(s).

Figures 8, 9, 10, and 11 show the evolution of the network over time. In 2000, very few statisticians started writing about biopharmaceutical statistics, the graph in Fig. 8 shows an isolated authors with two cliques of size three and two dyads. In Fig. 9, we start seeing more cliques, more groups are publishing in the biopharmaceutical subfield. In Fig. 10, the network is growing tremendously with more
[SAI2010, p.264]
individuals publishing, it seems like H. James and W. Jane are leading coauthors in the new field. Finally, in 2003, the subfield is well-established with several independent and mutually exclusive groups working simultaneously, the leading figures are still H. James and W. Jane.

“authors” is error introduced in editing.

[SHA2008, p.138, cont]
These attributes uniquely identify actors, some of which change rapidly/slowly over time while other attributes remain the same over time. For example, the attributes “name" and “field" do not change. The evolution of “weight" and “status" attributes can be viewed as a time series because they change faster than any other attributes. “Age" changes linearly over time. Meanwhile, the “active" flag operates as a switch initially set to “on", but later could change to “off”, once it is changed to “off" it remains in that state forever. Certain actors might change the attribute “preference".

The model was implemented in MatLab and consists of approximately 350 lines of code, it exploits the distributions of tie-strength and clique-size to build the coauthorship network. Figure 5.40(a) is a 2-mode author-by-paper simulated network.

Note that a new publication surfaces at each time step. Figure 5.40(b) shows the 1-mode coauthorship network corresponding to the matrix in Figure 5.40(a).

[SHA2008, p.141]
Figure 5.42 shows another simulated coauthorship social network, the program ran for 100 iterations. The simulated network is similar to the network obtained from empirical data, see section 5.3.3.

[SHA2008, p.137]
5.3.3 The Emergence of Scientific Subfields
The biopharmaceutical subfield joins the fields biology and pharmacy. In this part, I explore the biopharmaceutical statisticians socio-network over time to inspect the emergence of this discipline. The data include papers published between the years 1977 and 2003. There are 157 published papers with 260 unique coauthor(s).

Figure 5.38 shows the evolution of the network over time. In 2000, very few statisticians started writing about biopharmaceutical statistics, the graph in Figure 5.38(a) shows isolated authors with two cliques of size three and dyadic relations. In Figure 5.38(b), we start seeing more cliques, more groups are publishing in the biopharmaceutical subfield. In Figure 5.38(c), the network is growing tremendously with more

individuals publishing, it seems like H. James and W. Jane are leading coauthors in the new field. Finally, in 2003, the subfield is well-established with several independent and mutually exclusive groups working simultaneously, the leading Figures are still H. James and W. Jane.

[SAI2010, p.265]

6 The Network of Well-Established Scholars

Figure 1 presents the social network of prominent statisticians affiliated with US universities. In this section, we will use the method of deleting weak ties and pendants (nodes with degree = 1) to expose the important actors in the network. In coauthor social networks, weak ties and hanging nodes do not impose great impact on the status of the network, however, in other types of social networks weak ties could be crucial to the status and performance of the network. What is worth knowing in social networks is who maintains strong ties with who and who is connected

[SAI2010, p.266]

to the most actors, such authors resemble the heart of the network and their strong ties is the blood that keeps it alive and active.

To begin with,

brokerage roles are evident in this network. For example, the node “Lange N” in Fig. 1 can be in the cut-point set, this author is connected to four key player scholars in the network, namely, “Gelfand A”, “Carlin B”, “Wand M” and “Zeger S”. While maintaining good relations with prominent authors in the field of statistics, this author also connects structurally different parts of the network and styles of coauthorship.

In addition, “Louis T” can also be considered in the cut-point author set, he is in contact with two mutually exclusive subgroups of authors in which none of the members of each subgroup publishes with member(s) of the other subgroup. “Hall P”, “Diggle P” and “Gijbels I” are not cut-point authors but yet connected to key figures in the network, they are publishing with authors most of which are affiliated with different universities and geographically located in different continents. Further investigation reveals that some of these authors although they are not geographically in the same place, but they went to the same school, majored in the same field and spoke the same language and thus maintained good relations.

We proceed by first removing pendant authors (nodes with degree = 1) and then removing ties with weight = 1, Fig. 12 depicts the altered network. Thick edges indicate higher weight, the thicker the link is the higher the number of publications. Big nodes indicate higher degree, the bigger the node is the higher the number of coauthors that particular author has. The network is

not centric, in fact, it is more like a chain-network with network diameter = 12. It contains three separate components. In this layout, “Donoho” and “Gelfand” are far away from each other. However, “Zeger” and “Breslow” form two independent subnetworks. Finally, the author “Marron”, “Hall”, “Fan”, “Gijbels”, “Wand” and “Jones” are very close and similar authors, they form inbred subnetwork.

[SAI2010, p.267]

Figure 13 shows the network of authors having tie strength of seven or higher. Clearly, there are components of the original network consist of authors with high coauthored papers, members of each component form an elite group of well-trusted authors and coauthors.

[SHA2008, p.141]

5.3.5 The Network of Well-Established Scholars

Figures 5.32, 5.33 present the social network of prominent statisticians affiliated with US universities. I will use the method of deleting weak ties and ~~and~~ pendants ~~vertices~~ (vertices with degree = 1) to expose the important coauthors in the network. In coauthor social networks, weak edges and hanging vertices do not impose great impact on the status of the network, however, in other types of networks weak ties could be crucial to the status and performance of the network. What is worth knowing in social networks is who maintains strong ties with who and who is connected

to the most actors, such authors resemble the heart of the network and their strong ties is the blood that keeps it alive and active.

[SHA2008, p.142]

Brokerage roles are evident in this network. For example, the vertex “Lange N" in Figure 5.32 can be in the cut-point set, this author is connected to four key player scholars, namely, “Gelfand A", “Carlin B",”Wand M" and “Zeger S". While maintaining good relations with prominent authors in the field of statistics, this author also connects structurally different parts of the network and styles of coauthorship.

In addition, “Louis T" can also be considered in the cut-point author set, he is in contact with two mutually exclusive subgroups of authors in which none of the members of each subgroup publishes with member(s) of the other subgroup. “Hall P", “Diggle P" and “Gijbels I" are not cut-point authors but yet connected to key figures in the network, they are publishing with authors most of which are affiliated with different universities and geographically located in different continents. Further investigation reveals that some of these authors although they are not geographically in the same place, but they went to the same school, majored in the same field and spoke the same language and thus maintained good relations. .

Continuing with the same spirit,

I proceed by removing

ties with weight = 1, Figure 5.43 depicts the altered network. Thick edges indicate higher weight value, the thicker the link is the higher the number of publications. Big nodes indicate higher degree, the bigger the vertex is the higher the number of coauthors that particular author has. The network is

[SHA2008, p.143]

not centric, in fact, it is more like a chain-network with network diameter = 12. It contains three separate components. In this layout, “Donoho" and “Gelfand" are far away from each other. However, “Zeger" and “Breslow" form two independent subnetworks. Finally, the authors “Marron", “Hall”, “Fan", “Gijbels", “Wand" and “Jones" are very close and similar authors, they form inbred subnetwork.

Figure 5.44 shows the network of authors with tie strength of seven or higher. Clearly, there are components of the original network consist of authors with high coauthored papers. Members of each component form an elite group of well-trusted authors and coauthors.

[SAI2010, p.267, cont]

7 Conclusions

This work contained two parts; in part one, we used empirical data to in-

vestigate the distributions of tie-strength and clique-size in coauthorship social networks. The distribution of tie-strength among authors is a well-behaved power law; however, the distribution of clique size is lognormal.

In the second part, we developed a program to generate coauthorship networks based on the distributions of tie-strength and clique size.

The model takes into account the fact that authors/nodes status and attributes change over time. The resulting artificial network looked similar to a realworld social network in the Biopharmaceutical subfield.

Acknowledgements The work of Dr. Said is supported in part by Grant Number F32AA015876 from the National Institute on Alcohol Abuse and Alcoholism. The work of Dr. Wegman is supported in part by the Army Research Office under contract W911NF-04-1-0447. Both were also supported in part by the Army Research Laboratory under contract W911NF-07-1-0059. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Alcohol Abuse and Alcoholism or the National Institutes of Health.

References

Barabasi, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439),509–512. doi:10.1126/science.286.5439.509.

Borner, K., Dallasta, L., Ke, W., & Vespignani, A. (2005). *Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams*. Bloomington IN: Indiana University.

Carley, K. (2002). Smart agents and organizations of the future. In L. Lievrouw & S. Livingstone (Eds.), *The handbook of new media* (Chap. 12, pp. 206–220). Thousand Oaks, CA: Sage.

Cioffi-Revilla, C. (2005). *Power laws in the social sciences: Discovering complexity and nonequilibrium dynamics in the social universe*. Fairfax, VA: George Mason University.

Krackhardt, D., & Carley, K. (1998). PCANS model of structure in organizations. In *Proceedings of the 1998 international symposium on Command and Control Research and Technology* (pp. 113–119), Monterey, CA. Vienna, VA: Evidence Based Research.

Roth, C. (2005). Generalized preferential attachment: Towards realistic social network models. In *ISWC 4th intl Semantic Web Conference, Workshop on Semantic Network Analysis*, Galway, Ireland.

Said, Y., Wegman, E., Sharabati, W., & Rigsby, J. (2008). Social networks of author–coauthor relationships. *Computational Statistics and Data Analysis, 52*, 2177–2184. doi:10.1016/j.csda.2007.07.021.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. NewYork: Cambridge University Press.

[SHA2008, p.143, cont]

To sum up, this section addressed two issues, the first concerned empirical data to in-

[SHA2008, p.144]

vestigate the distributions of tie-strength and clique-size in coauthorship social networks. The distribution of tie-strength among authors is a well-behaved power law; however, the distribution of clique size is lognormal.

While the second concerned the development of a program to generate ~~random~~ coauthorship network,

the model takes into account the fact that authors' status and attributes change over time. The resulting artificial network looked similar to a real coauthorship social network ~~of statisticians~~ in the Biopharmaceutical subfield.

Comment: if the work above was Sharabati's alone, the acknowledgement at left is interesting. Did Wegman and Said submit this as part of their reports regarding those contracts?

All 8 references are in [SHA2008], although it is not itself referenced.

This the same as [SHA2008], Lucia Dall'Asta's name is misspelled.

A decent reference is:

Börner, K., Dall'Asta, L., Ke, W., & Vespignani, A. (2005). Studying the emerging global brain: Analyzing and visualizing the impact of coauthorship teams. *Complexity*, *10*(4), pp. 58-67.

This at least improved the reference. In [SHA2008], it was simply

[8] K. Carley, *Smart agents and organizations of the future*, Carnegie Mellon University.

This also offered a better reference. In [SHA2008], it was:

[34] D. Krackhardt and K. Carley, *Pcans model of structure in organizations*, Carnegie Mellon University.

This is the third improved reference. In [SHA2008], it was:

[51] C. Roth, *Generalized preferential attachment: Towards realistic social network models*, (2005).

Thanks to Ted Kirkpatrick and DC for comments, and *andrewt* for finding the new antecedents used to expand **W.5.7.**