Supporting Appendices

Table of Contents

Glossary

0047	W911NF-04-1-**0447** ARO-managed research grant, Wegman
0059	W911NF-07-1-**0059** ARO-managed research grant, Wegman
5876	F32AA01**5876** NIAAA alcoholism postdoc fellowship, Said
ack	acknowledgement in paper/talk, of people or especially funding
ARL	Army Research Laboratory,
ARO	Army Research Office, manages external research
claim	claim of paper, talk or other work in grant report to agency
DARPA	Defense Advanced Projects Agency
DC	Canadian blogger "Deep Climate" (person*), Deep Climate* (blog)
DHHS	US Department of Health and Human Services
DoD	US Department of Defense
FOIA	Freedom of Information Act
GMU	George Mason University, Fairfax, VA
IG	Inspector General (of DoD or DHHS, for example)
JHU	Johns Hopkins University, Baltimore, MD strong public health
NIAAA	National Institute on Alcohol Abuse and Alcoholism (in DHHS)
NIH	National Institutes of Health, of which NIAAA is one
NSWC	Naval Surface Warfare Center
ORI	Office of Research Integrity, research watchdog of DHHS[1]
SNA	Social Network Analysis, mis-applied in WR and [SAI2008]
WR	Wegman Report (2006),[2] [WEG2006], most of T126,

[1] ori.hhs.gov main ORI page
ori.hhs.gov/research-misconduct-0 research misconduct, especially plagiarism
ori.hhs.gov/case_summary 2011: Jagannathan, Lushington, Visvanathan, Weber
[2]archives.republicans.energycommerce.house.gov/108/home/07142006_Wegman_Report.pdf
[3]archives.republicans.energycommerce.house.gov/108/home/07142006_Wegman_fact_sheet.pdf This was the announcement.
[4] Joe Barton, et al, 07/19/26, 07/27/06. DC provides page-numbered PDF version: deepclimate.files.wordpress.com/2010/04/hockey-stick-hearings-2006-ec-committee.pdf The original, with unnumbered pages is at: frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=109_house_hearings&docid=f:31362.wais
[5]deepclimate.org/2009/12/17/wegman-report-revisited DC found first WR issues.
[6]deepclimate.org/2010/09/15/wegman-report-update-part-2-gmu-dissertation-review
[7] deepclimate.org/2010/11/16/replication-and-due-diligence-wegman-style
[8] www.desmogblog.com/crescendo-climategate-cacophony 185p
[9]www.desmogblog.com/sites/beta.desmogblog.com/files/STRANGE.SCHOLARSHIP.V1.02.pdf 250p originally hosted at Deep Climate, with many thanks. Now consolidated here.
[10] www.desmogblog.com/gmu-still-paralyzed-wegman-and-rapp-still-paranoid 45p
[11] www.desmogblog.com/mashey-report-reveals-wegman-manipulations 17p
[12] www.desmogblog.com/wegman-report-not-just-plagiarism-misrepresentation 12p
[13]www.desmogblog.com/curious-coincidences-george-mason-university-ed-wegman-milton-johns-and-ken-cuccinelli
[14] www.desmogblog.com/fake-science-fakexperts-funny-finances-free-tax 213p
[15]www.desmogblog.com/ed-wegman-promised-data-rep-henry-waxman-six-years-ago-where-it

Frequently-cited references

BAR2006	07/14/06	Report Raises New Questions About Climate Change Assessments, House Energy and Commerce Committee.[3]
BAR2006a	07/27/06	Complete Transcript, of Wegman Report[4]
DEE2009	12/17/09	Contrarian scholarship: Revisiting the Wegman Report[5]
DEE2010p	09/15/10	Wegman report update, part 2: GMU dissertation review[6]
DEE2010r	11/16/10	Replication and due diligence, Wegman style[7]
DEN2005	2005	Wouter de Nooy, Andrej Mrvar, Vladimir Batagelj, *Exploratory Social Network Analysis with Pajek (used to be online, not found)*
MAS2010	03/15/10	Crescendo to Climategate Cacophony[8]
MAS2010a	09/26/10	Strange Scholarship in the Wegman Report[9]
MAS2011	01/04/11	Strange Inquiries at George Mason University[10]
MAS2011a	05/26/11	Strange Tales and Emails – Said, Wegman, et al[11]
MAS2011b	05/27/11	Strange Falsifications in the Wegman Report[12]
MAS2011d	10/30/11	Curious coincidences at George Mason University …[13]
MAS2012	02/13/12	Fake Science, Fakexperts, Funny Finances, Free of tax[14]
MAS2012b	07/13/12	Ed Wegman Promised Data to Rep. Henry Waxman…[15]
MAS2012c	08/20/12	See No Evil, Speak Little Truth, Break Rules, …[16]

MAS2012d 10/25/12 Fakery 2: More Funny Finances, Free of Tax[17]
REZ2009 Spring'09 *Enhancement of Network Robustness and Efficiency through Evolutionary Computing, Statistical Computation and Social Network Analysis*[18] *Dissertation*
SHA2008 10/31/08 *Multi-Mode and Evolutionary Networks*[19]
SAI2007 09/07/07 *Experiences with Congressional Testimony: Statistics and The Hockey Stick, GMU Data and Statistical Sciences Colloquium*[20] T414
SAI2008 2008 Social networks of author–coauthor relationships,[21] P179
SAI2010 Yasmin H. Said, Edward J. Wegman, and Walid 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 2010, pp.257-268.[22] P200.
VER2010 10/08/10 University investigating prominent climate science critic[23]
VER2010a 11/22/10 Experts claim 2006 climate report plagiarized[24]
VER2010c 11/23/10 Wegman report round-up[25]
VER2011 05/15/11 Climate study gets pulled after charges of plagiarism[26]
VER2011a 05/16/11 Retracted climate critics' study panned by expert[27]
VER2012 02/22/12 Univ. reprimands climate science critic for plagiarism[28]
WAS1994 1994 Stanley Wasserman and Katherine Faust, *Social Network Analysis – Methods and Applications*, Cambridge
WEG2006 07/14/06 Wegman Report (**WR**) - Ad Hoc Committee Report On The 'Hockey Stick' Global Climate Reconstruction[29] T126 includes the WR and several days' testimony to Congress, including a follow-up report.
WEG2006b 07/27/06 Testimony of 07/27/06[30] This is also part of T126.
WEG2009 02/16/09 Proposal written for ARO, not accepted[31]
WEG2010 06/18/10 Resume of Edward Wegman, archived[32]
WEG2010a Nov 2010 Wegman FOIA "Relevant emails" to Dan Vergano[33]
Caveat: seems a partial collection, very relevant, but do not over-interpret.
WSJ2006 07/14/06 Hockey Stick Hokum[34] Wall Street Journal editorial, appeared same day as Barton announcement and contained a falsified image.[35]

Further reading

Raymond S. Bradley, *Global Warming and Political Intimidation*, 2011.
Michel E. Mann, *The Hockey Stick and the Climate Wars…*, 2012.

---

[16] www.desmogblog.com/see-no-evil-speak-little-truth-break-rules-blame-others
[17] www.desmogblog.com/2012/10/23/fakery-2-more-funny-finances-free-tax
[18] Hadi Rezazad, PhD Dissertation
gradworks.umi.com/33/64/3364566.html
[19] Walid Sharabati, Phd Dissertation
digilib.gmu.edu:8080/dspace/bitstream/1920/3384/1/Sharabati_Walid.pdf
[DEE2010p] and [MAS2012c §W.5.7] discuss the plagiarism problems.
[20] www.galaxy.gmu.edu/stats/colloquia/AbstractsFall2007/TalkSept7.pdf
Infinite thanks to DC for this, which revealed much hidden information.
This key file disappeared August 2010, but an annotated copy is included in [MAS2010a §A.11.2]. DC saved one also: www.webcitation.org/6E35F5rZr
deepclimate.files.wordpress.com/2010/09/said-talksept7.pdf
[21] Yasmin H. Said, Edward J. Wegman, Walid K. Sharabati, John T. Rigsby, "Social networks of author–coauthor relationships," *Computational Statistics & Data Analysis* 52 (2008) 2177 – 2184. Recvd 8 July 2007; accepted 14 July 2007. The (2007) vs (2008) difference has caused citation confusion. **Retracted**.
[22] link.springer.com/chapter/10.1007%2F978-3-642-03739-9_30?LI=true
[23] content.usatoday.com/communities/sciencefair/post/2010/10/wegman-plagiarism-investigation-/1 UPDATE 05/26/11 on Walsch comments
[24] www.usatoday.com/weather/climate/globalwarming/2010-11-21-climate-report-questioned_N.htm
[25] content.usatoday.com/communities/sciencefair/post/2010/11/wegman-report-round-up/1
[26] www.usatoday.com/weather/climate/globalwarming/2011-05-15-climate-study-plagiarism-Wegman_n.htm
[27] content.usatoday.com/communities/sciencefair/post/2011/05/retracted-climate-critics-study-panned-by-expert-/1
[28] content.usatoday.com/communities/sciencefair/post/2012/02/george-mason-university-reprimands-edward-wegmand-/1
[29] archives.republicans.energycommerce.house.gov/108/home/07142006_Wegman_Report.pdf Edward J. Wegman, David W. Scott, Yasmin H. Said
[30] archives.republicans.energycommerce.house.gov/108/Hearings/07272006hearing2001/Wegman.pdf part of [BAR2006]
[31] www.documentcloud.org/documents/524550-descriptf-network-science-proposal.html
[32] web.archive.org/web/20100609135746/www.galaxy.gmu.edu/stats/faculty/wegman.resume2.pdf
www.galaxy.gmu.edu/stats/faculty/wegman.resume2.pdf originally, but it and other key files vanished [MAS2010, §A.11.1], ~between 08/16/10 and 08/23/10.
[33] www.documentcloud.org/documents/527437-relevant-emails-redact-addresses.html
[34] archives.republicans.energycommerce.house.gov/108/News/07142006_1990.htm
[35] sg.wsj.net/public/resources/images/ED-AE505_1hocke_20060713182815.gif
The false citation's odd history is explained:
scienceblogs.com/stoat/2012/10/08/more-use-and-abuse-of-ipcc-1990-fig-7-1c

Key people,
Most are discussed in [MAS2010a].
Stanley Azen, USC, past Editor-in-Chief of *CSDA* [MAS2011a]
Joseph Barton (R-TX), **Ed Whitfield** (R-KY), US Representatives who got Wegman recruited, and promoted the WR
Jerry Coffey, associate of Jim Tozzi (Data Quality Act), consultant for House Republicans, old associate of Wegman's, contacted him for WR
Myron Ebell, CEI, Cooler Heads Coalition, with Fred Singer, one of two key recruiters of McKitrick and McIntyre, [MAS2010a]
Milton Johns, lawyer for Wegman and Said, [MAS2011d].
Steven McIntyre, retired mining consultant, Ontario, Canada. With McKitrick, created talk that acted as WR blueprint [MAS2011a, p.17].
Ross McKitrick, Prof. Economics, U of Guelph, Ontario, Canada
Pat Michaels, was at U VA, now CATO. In 2010, was a GMU Distinguished Senior Fellow, taught course for School of Public Policy.
Fred Singer, SEPP [MAS2012] long affiliated with GMU's Institute for Humane Studies in the 1990s, worked closely with GMI.
Peter Spencer, Barton staffer, met Wegman after Coffey, sent materials

Contributors to WR and related work (Wegman, associates)
Edward J. Wegman, GMU
David W. Scott, Rice University
Yasmin H. Said, PhD 2005 (Wegman), Johns Hopkins University (2005-2006), then back at GMU by date of WR release.
An unknown 4th person, who later dropped out[36]
WR ack'd 2 Wegman students for help, vaguely:
John T. Rigsby III, Naval Surface Warfare Center, MS 2005
Denise M. Reeves, MITRE, PhD 2009, wrote the oft-copied SNA text, , clarified later by Wegman [MAS2011a]

Walid Sharabati, PhD, 2008. Unmentioned in the WR, he contributed much of response to Rep. Stupak in 2006, [WEG2006b], part of T126.

[36] *It might be worthwhile to locate this person and ask them to comment.*

X. Plagiarism in another 5876-claimed paper

Said claimed P405 S③☆ for 5876 in **§S.3.2**[37]:

'20. Said, Yasmin H. (2007) On the Eras in the History of Statistics and Data Analysis, *Journal of Washington Academy of Sciences*, 93(1), 17-35'[38]

This was 5876-unfit, but also *seemed* an odd effort for a busy postdoc, less than 2 years after PhD. The paper started, p.17:

'Yasmin H. Said
Center for Computational Statistics
George Mason University
Abstract
In this paper, **we**[39] present a view of the evolution of statistical thinking through eras we designate as Pre-modern, Classical, Recent Past, and Future. **We** argue that modes of thinking about data and statistical inference are noticeably different from one era to the next. **We** discuss some of the leading figures in each of these eras.'

The paper, p.18 describes this as:

'Pre-modern Period prior to 1900
(pp.18-19, many)
Classical Period 1900 to 1985
(pp.19-24,K. Pearson, Gossett, Fisher, E.Pearson, Neyman, Kolmogorov, Mahalanobi, Hotelling, Cramér, Rao, Wilks)
Recent Past Period 1962 to 2005
(pp.25-26, Tukey and colleagues)
Future Period after 1981.'
(pp. 26-30, Wegman, Efron, Friedman, David Scott)
Statistical Thinking in Government, Science, and Law, Conclusions, etc
(pp. 3135-, Wegman, Efron, Friedman, David Scott)

A few hours were spent documenting obvious examples of mosaic plagiarism, with no pretense of thorough search. Similar phrases appear in many Internet sources, so one cannot be sure of the adaptation flow, just the existence of at least one plausible prior antecedent. However, it is *almost certain* that such existed in various websites before this paper.

[37] www.desmogblog.com/sites/beta.desmogblog.com/files/aa15876-3-3-Progress.Redact.pdf , PDF p.5.
[38] www.washacadsci.org/scans/V.93-n.1.pdf
The article PDF itself was not online, but the Washington Academy of Sciences kindly and promptly sent me a copy.
[39] The "we" usage might be stylistic or it might imply this was originally planned as a paper by Said and Wegman.

The side-by-side comparison style is the same as used in earlier reports:

- Cyan for identical, in-order text between P405 and antecedents, with some reformatting for alignment.
- Yellow for trivial edits P405 ← antecedents.
- ~~Cross out~~ is used for obviously-deleted antecedent text.
- No highlight implies plausible paraphrase, original work, or unfound.

Said wrote, p.13:

'Acknowledgement
The author gratefully acknowledges the long discussions with Professor Edward J. Wegman, whose contact and experience with both the early contributors and the evolution of statistics as a discipline over the last 40 years provided valuable insight that made this discussion possible.'

Substantial text of "striking similarity" was quickly identified, in the same copy-paste-edit style of which 90+ pages have been documented so far.

P405. ← Plausible antecedent

Page	antecedent source
18	Stephen M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900*, Belknap Press of Harvard University Press (March 1, 1990)[40]
20	www-history.mcs.st-andrews.ac.uk/Biographies/Pearson.html
20-21	www-gap.dcs.st-and.ac.uk/~history/Biographies/Fisher.html
22	www-history.mcs.st-andrews.ac.uk/Biographies/Pearson_Egon.html
22	statprob.com/encyclopedia/PrasantaChandraMAHALANOBIS.html
22	www.amstat.org/about/statisticiansinhistory/index.cfm?fuseaction=biosinfo&BioID=7
23	www-history.mcs.st-andrews.ac.uk/Biographies/Cramer_Harald.html
24	www-history.mcs.st-andrews.ac.uk/Biographies/Wilks.html
25	**P169** Wegman, Edward J. and Solka, Jeffrey L.(2005) "Statistical data mining," *Handbook of Statistics: Data Mining and Data Visualization* (Rao, C. R., Wegman, E. J. and Solka, J. L., eds.), 1-46'[41]

This paper has 17 pages of text, plus 2 of endnotes and references. A few hours' work easily found ~6 pages of text copy-paste-edited from unacknowledged sources, casting doubt on Said's qualification to write a statistics history article for an encyclopedia, i.e., P403, **§Y**.

[40] The hardback was (1986), paperback (1990). This seems more paraphrased than the other copy-paste sections.. Stigler's book was well-known, heavily-cited, and 2 decades old, so intermediate antecedents were easily possible.
[41] That book also contained P402 S☆.

P405 p.18

In the Pre-modern Period, one of the most interesting early examples of the recognition of variability is the so-called Trial of the Pyx. The Trial of the Pyx is a procedure for maintaining the integrity of newly minted coins in the United Kingdom (England). From shortly after the Norman Conquest (1066) in a procedure that has been essentially unchanged since 1282, the London (later Royal) Mint

selects a sample of each day's coins that are reserved in a box called the Pyx. The earliest agreements between the mint and the monarchy stated that a certain tolerance would be allowed in the weight of a single coin and by linear extrapolation in the aggregate weight of the contents of the Pyx. Thus, earlier than 1100, there was a formalized methodology for allowance of uncertainty and a method by which the integrity of the entire coinage could be judged based on a sample in the presence of uncertainty in the production process[i]

--- (note on p.33)

[i] The use of linear extrapolation is a flawed procedure by modern standards. If a tolerance of 2 units per coin is allowed, then for 100 coins, the Trial of the Pyx would allow 200 units tolerance, whereas modern theory would dictate a tolerance of 2√100 =20 units tolerance.

P405 p.20 (on Pearson)

lectures on statistics,[42] dynamics and mechanics, completed the unfinished first volume of Clifford's *The Common Sense of the Exact Sciences* (published in 1885), completed and edited the half-written first volume of Todhunter's History of the *Theory of Elasticity*, began working on the second volume published many papers on applied mathematics, lectured on *The Ethic of Free Thought*, and undertook research on a number of historical topics, including the evolution of Western Christianity. …

Stephen M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900*, Belknap Press of Harvard University Press (March 1, 1990)[43]
p.3

One dramatic early instance of a numerical assessment of accuracy that was not given in terms of explicit probabilities was the Trial of the Pyx.

From shortly after the Norman Conquest up to the present, the London (later Royal) Mint maintained the integrity of its coinage[44] through a routinized inspection scheme in which a selection of each day's coins was reserved in a box ("the Pyx") for a later trial. Even in the earliest indentures between the mint and the king ~~the contract~~ stated that the trial would allow a tolerance in the weight of a single coin and, by linear extrapolation, in the aggregate weight of the entire contents of the Pyx. Thus as early as 1100 an economic necessity had led to an institutionalized numerical allowance for uncertainty, uncertainty in how the value of the entire coinage could be judged by that sample, in the presence of unavoidable variability in the production process.[2]

---

2. The Trial of the Pyx was not without its flaws. The use of linear extrapolation was a major one. If a coin was allowed a tolerance of 5 grains, an aggregate of 100 coins would be allowed a tolerance of 500 grains, rather than the √100 X5 = 50 grains modern theory might suggest. The story of the Pyx, including Isaac Newton's connection to it, is told in Stigler(1977b).

www-history.mcs.st-andrews.ac.uk/Biographies/Pearson.html

lectures on statics, dynamics and mechanics, ~~he~~ completed the unfinished first volume of Clifford's The Common Sense of the Exact Sciences (published in 1885), completed and edited the half written first volume of Todhunter's History of the Theory of Elasticity, began working on the second volume ~~which had hardly been started by Todhunter, and~~ published many papers on applied mathematics. ~~He also~~ lectured on The Ethic of Free Thought, and undertook research on a number of historical topics such as the evolution of Western Christianity. …

[42] Google: karl pearson lecture statics dynamics
Statics is a specific topic within physics, as per "Statics of Rigid Bodies" as in Chapter 14 in my sophomore college physics book Halliday and Resnick (1963). **Statics** is very definitely not **statistics**, but this sort of error has been found often in works involving Said, including P401 and T126: amidst a block of obviously-copied text, trivial edits were made, but some introduced errors. Some were really silly, indicating lack of understanding that went far beyond poor proofreading.

[43] www.amazon.com/The-History-Statistics-Measurement-Uncertainty/dp/067440341X

[44] This likely got rearranged into "maintaining the integrity of newly minted coins," but the (manual) comparison algorithm does not try to track movements of text, i.e., it approximates the old UNIX *diff(1)* command.

P405 p.20 (cont)

'Sir Ronald Fisher (1890-1962) is widely recognized as the third and probably most important of the first modern statisticians. He studied mathematics and astronomy at Cambridge, but was also interested in biology. He graduated with distinction in the Mathematical tripos of 1912. He continued his studies at Cambridge on the theory of errors. Fisher's interest in the theory of errors eventually led him to investigate statistical problems. After leaving Cambridge, Fisher worked for several months on a farm in Canada, but soon returned to London and took up a position as a statistician in the Mercantile and General Investment Company. When war

P405 p.21

broke out in 1914 he tried to enlist in the army, having already trained in the Officers' Training Corps while at Cambridge. He was rejected for military service because of his eyesight. He became a teacher of mathematics and physics, teaching at Rugby and other similar schools between 1915 and 1919. Fisher gave up being a mathematics teacher in 1919 when he was offered two posts simultaneously. Karl Pearson offered him the post of chief statistician at the Galton laboratories, but he was also offered the post of statistician at the Rothamsted Agricultural Experiment Station, which was the oldest agricultural research institute in the United Kingdom. It was established in 1837 to study the effects of nutrition and soil types on plant fertility, and this appealed to Fisher's interest in farming. He accepted the post at Rothamsted. Here he made many contributions to statistics, in particular the design and analysis of experiments, and also to genetics. He studied the design of experiments by introducing the concept of randomization and the analysis of variance, procedures now used throughout the world. In 1921 he introduced the concept of likelihood. The likelihood of a parameter is proportional to the probability of the data, and it gives a function that usually has a single maximum value, which he called the maximum likelihood. Fisher published a number of important texts; in particular, *Statistical Methods for Research Workers* (1925) ran to many editions that he extended throughout his life.

Pearson and Fisher had a long, bitter, and very public dispute. At first they exchanged friendly letters after Pearson received a manuscript from Fisher in September 1914 of a paper he was submitting for publication to *Biometrika*. Pearson's initial response was to offer his hearty congratulations and, if correct, offered to publish the paper Later, having read the paper fully he indicated that it marked a distinct advance.

By May 1916 they were still corresponding in a friendly manner. However, Pearson misunderstood the assumptions of Fisher' maximum likelihood method, and criticized it in his May 1917 *Cooperative Study,* a paper that he co-authored with his staff concerning tabulating the frequency curves. Fisher, believing that Pearson's criticism was unwarranted, responded with a paper that criticized examples in the *Cooperative Study* to the extent of ridiculing them. Fisher had looked again at his earlier correspondence with Pearson, noticed that many of his papers had been rejected, and concluded that Pearson had been responsible. Thus began one of the most famous feuds in the history of statistics.'

www-gap.dcs.st-and.ac.uk/~history/Biographies/Fisher.html

'Although he studied mathematics and astronomy at Cambridge, he was also interested in biology. … He graduated with distinction in the mathematical tripos of 1912. Awarded a Wollaston studentship, he continued his studies at Cambridge under Stratton on the theory of errors reading Airy's manual the *Theory of Errors.* It was Fisher's interest in the theory of errors that eventually led him to investigate statistical problems. After leaving Cambridge, Fisher had no means of financial support and worked for a few months on a farm in Canada. He returned to London, taking up a post as a statistician in the Mercantile and General Investment Company. When war

broke out in 1914 he enthusiastically tried to enlist in the army, having already trained in the Officers' Training Corps while at Cambridge. His medical test showed him A1 on all aspects except his eyesight, which was rated C5, so he was rejected. He became a teacher of mathematics and physics, teaching at Rugby and other similar schools between 1915 and 1919. Fisher gave up being a mathematics teacher in 1919 when he was offered two posts simultaneously. Karl Pearson offered him the post of chief statistician at the Galton laboratories and he was also offered the post of statistician at the Rothamsted Agricultural Experiment Station. This was the oldest agricultural research institute in the United Kingdom, established in 1837 to study the effect of nutrition and soil types on plant fertility, and it appealed to Fisher's interest in farming. He accepted the post at Rothamsted where he made many contributions ~~both~~ to statistics, in particular the design and analysis of experiments, and to genetics. ~~There~~ he studied the design of experiments by introducing the concept of randomisation and the analysis of variance, procedures now used throughout the world. … In 1921 he introduced the concept of likelihood. The likelihood of a parameter is proportional to the probability of the data and it gives a function which usually has a single maximum value, which he called the maximum likelihood … Fisher published a number of important texts; in particular *Statistical Methods for Research Workers* (1925) ran to many editions which he extended throughout his life.

www.educ.fc.ul.pt/icm/icm2003/icm14/Pearson.htm

Pearson had a long, bitter, and very public dispute with Fisher. At first they exchanged friendly letters after Pearson received a manuscript from Fisher in September 1914 of a paper he was submitting for publication. Pearson's initial response was to say (see [18]): *I congratulate you very heartily on getting out the actual distribution form ... if the analysis is correct which seems highly probable, I should be delighted to publish the paper in Biometrika.* Again a week later [18]:- *I have now read your paper fully and think it marks a distinct advance.* … By May 1916 they were still corresponding in a friendly manner. However Pearson misunderstood the assumptions of Fisher's maximum likelihood method, and criticised it unfairly in the May 1917 *Cooperative Study* paper which he co-authored with his staff concerning tabulating the frequency curves. Fisher, believing that Pearson's criticism was unwarranted, responded with a paper which criticised examples in the *Cooperative Study* to the extent of ridiculing them. Fisher had looked again at his earlier correspondence with Pearson, noticed that many of his papers had been rejected, and concluded that Pearson had been responsible.'

P405 p.22

'Egon Pearson (1895-1980) was the son of Karl Pearson
In 1921 he joined his father's Department of Applied Statistics at University College London as a lecturer. However, his father kept him away from lecturing. Egon attended his father's lectures and began to produce a stream of high quality research publications on statistics. In 1924, Egon became an assistant editor of *Biometrika*, but perhaps one of the most important events for his future research happened in the following year.

Jerzy Neyman (1894-1981) was stimulated by a letter from Egon Pearson, who sought a general principle from which Gosset's tests could be derived. Neyman went on to produce fundamental results on hypothesis testing and, when Egon Pearson visited Paris in the spring of 1927, they collaborated in writing their first paper
Between 1928 and 1933, they wrote a number of fundamental papers on hypothesis testing, the best-known result being the Neyman-Pearson Lemma. Neyman moved to the University of California, Berkeley in 1938 and remained there until his death in 1981. He was reputed to have been working on a research paper in the hospital where he died.'

Andrei Nikolaevich Kolmogorov (1903-1987) laid the axiomatic foundations for probability theory in 1933 and also in 1938 laid out the foundations for Markov random processes.

Prasanta Chandra Mahalanobis (1893-1972) undertook work on experimental designs in agriculture. In 1924, he made some important discoveries about the probable error of results of agricultural experiments, which put him in touch with Fisher. Later in 1926, he met Fisher at the Rothamsted Experimental Station and a close personal relationship was immediately established that lasted until Fisher's death.
In 1927, Mahalanobis spent a few months in Karl Pearson's laboratory in London. During this period he performed extensive statistical analyses of anthropometric data and closely examined Pearson's Coefficient of Racial Likeness (CRL) for measurement of biological affinities. He noted several shortcomings of the CRL and in 1930 published his seminal paper on the *D-square* statistic, which is now recognized as the Mahalanobis Distance.'

www-history.mcs.st-andrews.ac.uk/Biographies/Pearson_Egon.html

'In 1921 Pearson joined his father's Department of Applied Statistics at University College London as a lecturer. However, ~~despite being a lecturer~~, his father ~~seems to have~~ kept him away from lecturing. ~~Instead~~ Pearson attended ~~all of~~ his father's lectures and began to produce a stream of high quality research publications on statistics. In 1924 Pearson became an assistant editor of *Biometrika* but perhaps one of the most important events for his future research happened in the following year.
www-history.mcs.st-and.ac.uk/Biographies/Neyman.html
However his interest in statistics was stimulated ~~again~~ by a letter from Egon Pearson, who sought a general principle from which Gosset's tests could be derived. Neyman went on to produce fundamental results on hypothesis testing and, when Egon Pearson visited Paris in the spring of 1927, they collaborated in writing their first paper. …
Between 1928 and 1933 Neyman and Egon Pearson had written a number of important papers on hypothesis testing …'

statprob.com/encyclopedia/PrasantaChandraMAHALANOBIS.html

Prasanta Chandra MAHALANOBIS b. 29 June 1893 - d. 28 June 1972
Some of the early statistical studies he undertook were on experimental designs in agriculture. In 1924, he made some important discoveries pertaining to the probable error of results of agricultural experiments, which put him in touch with R.A. Fisher (q.v.). Later in 1926, he met Fisher at the Rothamsted Experimental Station and a close personal relationship was immediately established which lasted until Fisher's death.
In 1927, Mahalanobis spent a few months in Karl Pearson's (q.v.) laboratory in London, during which period he performed extensive statistical analyses of anthropometric data and closely examined Pearson's Coefficient of Racial Likeness (CRL) for measurement of biological affinities. He noted several shortcomings of the CRL and in 1930 published his seminal paper on the D^2-statistic entitled ``Tests and measures of group divergence".'

P405 p.22

'Harold Hotelling (1895-1973) earned a Ph.D. in mathematics from Princeton University, and began teaching at Stanford University that same year, 1924. Hotelling realized that the field of statistics would be more useful if it employed methods of higher mathematics, so in 1929, he went

P405 p.23

to England to study with R. A. Fisher.
When Hotelling returned to the United States, he began developing some of his techniques at Stanford University. His early applications involved the diverse fields of journalism, political science, population, and food supply.
Hotelling was a pioneer in the field of mathematical statistics and economics in the 20th century, with contributions to the theory of demand and utility, welfare economics, competition, game theory, depreciation, resource exhaustion, and taxation. His work in mathematical statistics included his famous 1931 paper on the Student's *t*-distribution for hypothesis testing, in which he laid out what has since been called *confidence intervals*.

P405 p.23

Carl Harald Cramér (1893-1985) entered the University of Stockholm in 1912 and worked as a research assistant on a biochemistry project before becoming firmly settled on research in mathematics.
He earned a Ph.D. in 1917 for his thesis, *On a class of Dirichlet series*.
In 1919 Cramér was appointed assistant professor at the University of Stockholm. He began to produce a series of papers on analytic number theory. It was through his work on number theory that Cramér was led towards probability theory. He also had a second job, namely as an actuary with the Svenska Life Assurance Company. This led him to study probability and statistics that then became the main area of his research. Cramér became interested in the rigorous mathematical formulation of probability in work of the French and Russian mathematicians, in particular the axiomatic approach of Kolmogorov. By the mid 1930s Cramér's attention had turned to the approach of the English statisticians such as Fisher and Egon Pearson as well as contemporary American statisticians.
During World War II, Cramér was cut off from the rest of the academic world. By the end of World War II he had written his masterpiece *Mathematical Methods of Statistics*. In addition to his seminal book, Cramér is known for his work on stationary stochastic processes and for the Cramér-Rao inequality.'

www.amstat.org/about/statisticiansinhistory/index.cfm?fuseaction=biosinfo&BioID=7

Harold Hotelling 1895-1973 … In 1924, he earned a PhD in mathematics from Princeton University, and began teaching at Stanford University that same year. Hotelling soon realized that the field of statistics would be more useful if it employed methods of higher mathematics, so in 1929, he went

to England to study with R. A. Fisher, a very prominent statistician.
When Hotelling returned to the United States, he began developing some of his techniques at Stanford University. His early applications involved the diverse fields of journalism, political science, population, and food supply. …
Hotelling was considered a pioneer in the field of mathematical statistics and economics in the 20th century, with contributions to the theory of demand and utility, welfare economics,
and taxation. His work in mathematical statistics included his famous 1931 paper on the Student's t distribution for hypothesis testing, in which he laid out what has since been called "confidence intervals." His economics papers throughout the 1920s and 1930s discussed
competition, game-theory, depreciation, and resource exhaustion. He also covered topics in mathematical statistics such as hypothesis testing and confidence intervals.'

www-history.mcs.st-andrews.ac.uk/Biographies/Cramer_Harald.html

Harald Cramér entered the University of Stockholm in 1912. …
worked as a research assistant on a biochemistry project before becoming firmly settled on research in mathematics. …
resulted in the award of a PhD in 1917 for his thesis *On a class of Dirichlet series*.
In 1919 Cramér was appointed assistant professor at the University of Stockholm. He began to produce a series of papers on analytic number theory ,… It was not only through his work on number theory that Cramér was led towards probability theory . He also had a second job, namely as an actuary with the Svenska Life Assurance Company. This led him to study probability and statistics which then became the main area of his research. Cramér became interested in the rigorous mathematical formulation of probability in work of the French and Russian mathematicians ~~such as Paul Lévy , Sergei Bernstein , and Aleksandr Khinchin in the early 1930s, but~~ in particular the axiomatic approach of Kolmogorov . By the mid 1930s Cramér's attention had turned to look at the approach of the English and American statisticians such as Fisher , ~~Neyman~~ and Egon Pearson ~~(Karl Pearson 's son)~~ …
During World War II Cramér was ~~to some extent~~ cut off from the rest of the academic world. … By the end of World War II Cramér had written his masterpiece *Mathematical Methods of Statistics*'.

P405 p.24

'Samuel Wilks (1906-1964)
began to study mathematics at the University of Texas in 1926 where he was taught set theory and other courses in advanced mathematics.
Wilks received an M.A. in mathematics in 1928.

Wilks was awarded a fellowship to the University of Iowa where he studied for his doctorate under H. L. Rietz. Rietz introduced him to Gosset's theory of small samples and R. A. Fisher's statistical methods. After receiving his doctorate in 1931 on small sample theory of 'matched' groups in educational psychology, he continued research at Columbia University in the 1931-1932 session.
In 1932, Wilks spent a period in Karl Pearson's department in University College, London. In 1933 he went to Cambridge where he worked with John Wishart, who had been a research assistant to both Pearson and Fisher. He was appointed instructor of mathematics at Princeton in 1933. He was to remain there for the rest of his career, being promoted to professor of mathematical statistics in 1944.
Wilks's work was all on mathematical statistics. His early papers on multivariate analysis were his most important, one of the most influential being, *Certain generalizations in the analysis of variance*. He constructed multivariate generalizations of the correlation ratio and the coefficient of multiple correlation and studied random samples from a normal multivariate population.
He advanced the work of Neyman on the theory of confidence-interval estimation. In 1941, Wilks developed his theory of 'tolerance limits.' Wilks was a founder member of the Institute of Mathematical Statistics (1935). There are obviously many other important contributors to the development of statistical theory in this Classical Period, but the ones mentioned here will suffice to give a flavor of the group. Much theory and methodology in the sense of the Classical Period still continues to be developed.

P405 p.25

In the landmark 1962 paper of Tukey entitled, "The future of data analysis," and later in the 1977 book, *Exploratory Data Analysis*,[iii] Tukey sets forth a new paradigm for statistical analysis. In contrast to confirmatory analysis in which a statistical model is assumed and inference is made on the parameters of that model, exploratory data analysis (EDA) is predicated on the fact that one does not

necessarily know that model assumptions actually hold for data under investigation. Because the data may not conform to the assumptions of the confirmatory analysis, inferences made with invalid model assumptions are subject to (potentially gross) errors. The idea then is to explore the data to verify that the model assumptions actually hold for the data in hand.'

P405 p.33

'The author gratefully acknowledges the long discussions with Professor Edward J. Wegman, whose contact and experience with both the early contributors and the evolution of statistics as a discipline over the last 40 years provided valuable insight that made this discussion possible.'

www-history.mcs.st-andrews.ac.uk/Biographies/Wilks.html

'During session 1926-27 Wilks taught at a school in Austin, Texas and at the same time he began to study mathematics at the University of Texas. Here he was taught set theory and other courses in advanced mathematics by Robert Moore and he took courses in probability and statistics with E L Dodd. Wilks received an M.A. in mathematics in 1928 and during this time, in fact from 1927 until 1929, he was an instructor in mathematics.
Wilks was awarded a fellowship to the University of Iowa where he studied for his doctorate. Here H L Rietz, who supervised his doctorate, introduced him to Gosset's theory of small samples and R A Fisher's statistical methods. After receiving his doctorate in 1931, on small sample theory of 'matched' groups in educational psychology, he continued research at Columbia University in session 1931-32.
In 1932 Wilks ~~went to England where he~~ spent a period in Karl Pearson's department in University College, London. In 1933 he went to Cambridge where he worked with John Wishart, who had been a research assistant to both Pearson and Fisher. He was appointed instructor of mathematics at Princeton in 1933. He was to remain there for the rest of his career, being promoted to professor of mathematical statistics in 1944.
Wilks's work was all on mathematical statistics. His early papers on multivariate analysis were his most important, one of most influential being *Certain generalizations in the analysis of variance*. He constructed multivariate generalisations of the correlation ratio and the coefficient of multiple correlation and studied random samples from a normal multivariate population. …
He advanced the work of Neyman on the theory of confidence-interval estimation. In 1941 Wilks developed his theory of 'tolerance limits'. Wilks was a founder member of the Institute of Mathematical Statistics (1935). ….

P169[45] Wegman and Solka (2005) p.2

The landmark paper of Tukey (1962) entitled, "The future of data analysis," and later in the book, *Exploratory Data Analysis* John Tukey(1977) sets forth a new paradigm for statistical analysis. In contrast to confirmatory analysis in which a statistical model is assumed and inference is made on the parameters of that model, exploratory data analysis (EDA) is predicated on the fact that we do not

p.3

necessarily know that ~~the~~ model assumptions actually hold for data under investigation. Because the data may not conform to the assumptions of the confirmatory analysis, inferences made with invalid model assumptions are subject to (potentially gross) errors. The idea then is to explore the data to verify that the model assumptions actually hold for the data in hand.'

[45] www.amazon.com/s/ref=nb_sb_noss?url=search-alias%3Daps&field-keywords=handbook%20of%20statistics%20wegman%20solka

Y. "Statistics" in Encyclopedia, possible plagiarism

P403 was a 2-page article listed by Said in support of her 5876 proposal:[46]

'2006 "Statistics" to appear *Encyclopedia of the Modern World*, (Stearns, Peter N., ed.), New York: Oxford University Press.

Externally-visible chronology was:

2005.08	Said gave 2006 as publication for *Encyclopedia, P403* .
2007.spring	19-page P405was published, **§X.**
2008.03	*Encyclopedia*[47] was actually published (Mar 28, 2008)

Given the inherent long creation time for the 8-volume *Encyclopedia*, ordering is hard to know, but 3 alternatives seem possible:

- *P403 might have been written first, then expanded into P405.*
- *P405 might have already existed in 2005, and then been edited to P403*
- *Draft P405 was written in 2006/2007, either by Wegman and Said or by Said drawing heavily on Wegman knowledge. Said then extracted a shorter version for the Encyclopedia, where it was further edited.*

The last alternative *seems likeliest*, from the evidence:

- P403's structure defines 4 historical periods and covers each. P405 retains some of this structure, but *oddly* devotes paragraphs to material long before the 1750-present period of the *Encyclopedia*.
- P403 is more crisply copy edited than P405.
- P403 retains odd vestiges of P405, but also cited a 2006 book not mentioned by it, making P403 unlikely to have been done in 2005.
- Stearns likely asked Wegman, and obvious choice. *He may have suggested Said, who could then could add P403 as further support for 5876. I'd defer serious assessment to statisticians, but P403 seemed strange* - **the only 20th-century statistician discussed was Wegman.**

It may have been misleading for Said to list this in support for a grant proposal and it may have been a distraction, but plagiarism is less clear. P403 was clearly derived from P405, but most of the obvious P405 plagiarism text was deleted, except the "Pyx" discussion.

Copyedits make more sense in this direction, English improved.

P403 p.135

'**STATISTICS is** at once an academic discipline, a tool for analyzing and inferring conclusions from data, and a collection subjected to the application of statistical tools. Statisticians generally think of the word statistics as ~~referring~~ either ~~to~~ the discipline or ~~to~~ the body of statistical methods whereas the general public more often thinks of statistics in the third sense, as a collection of numerical data, as in 'sports statistics.' The word "statistics" has its origins in the Latin *statisticum collegium* meaning "council of state." Similarly, the Italian word *statista* means "statesman" or "politician." Generically, statistics refers to data about the state. The modern English term derives from the German word *Statistik,* popularized and perhaps coined by the political scientist Gottfried Achenwall (1719-1772) in his *Vorbereitung zur Staatswissenschaft* (1748). The word seems to have been introduced into the English language by Sir John Sinclair (1754-1835). Sinclair was the supervisor of the twenty-one-volume *Statistical Account of Scotland,* published in the 1790s, which was the first systematic attempt to compile social and economic data on every parish in the country. In the *Statistical Account of Scotland*, Sinclair describes where he had come across the word *statistics* and why he translated and used it as an English word.'

P405 p.17

'Abstract

In this paper, we present a view of the evolution of statistical thinking through eras we designate as Pre-modern, Classical, Recent Past, and Future. We argue that modes of thinking about data and statistical inference are noticeably different from one era to the next. We discuss some of the leading figures in each of these eras.'

THE WORD "STATISTICS" refers at once to an academic discipline, to a powerful tool for inference on data and to results of the collection and application of statistical tools to data. Statisticians generally think of the word statistics as either the discipline or the body of methods comprising the tool while the general public more often thinks of statistics in the third sense, that is, a collection of numerical data as in 'sports statistics.' The word *statistics* is derived from the Latin *statisticum collegium* meaning the *council of state.* Similarly, the Italian word *statista* means *statesman* or *politician.* Thus, generically statistics means data about the state. The more modern term seems to have been the German word *Statistik,* popularized and perhaps coined by the German political scientist Gottfried Achenwall (1719-1772) in his *Vorbereitung zur Staatswissenschaft* (1748). The word *statistics* seems to have been introduced as an English language word by Sir John Sinclair (1754-1835). Sinclair was the supervisor of the *Statistical Account of Scotland* (1791- 1799), which was published in 21 volumes and was the first systematic attempt to compile social and economic statistics for every parish in the country. In the *Statistical Account of Scotland*, Sinclair describes where he had come across the word *statistics* and why he translated and used it as an English word.'

[46] www.desmogblog.com/sites/beta.desmogblog.com/files/aa15876-1a1-1-Proposal.Redact_0.pdf p.6

[47] www.amazon.com/Oxford-Encyclopedia-Modern-World-Present/dp/B007MXUUSU/ref=sr_1_1?ie=UTF8&qid=1360035163.

P403 p.135 (cont)

'The development of statistics as an academic discipline parallel the ever-increasing amounts of data generated by states and institutions.

The first U.S. Census was taken under the authority of Secretary of State Thomas Jefferson in 1790, when U. S. Marshals on horseback counted 3.9 million people. By 1810, the U. S. census was expanded to obtain information on manufacturing, including the amount and value of products. By 1840,

the American Statistical Association had been founded.

In the United States,[48] Abraham Lincoln established the United States Department of Agriculture (USDA) in 1862 to collect and analyze information pertaining to the country's agrarian economy.

A major development took place in Europe in 1953 with the development of the European Statistical System

(Eurostat), which for'

P403 p.136 (cont)

'the first time, integrated statistics across all of Western Europe. The discipline's close association with the state continues to facilitate advances in survey research and sampling theory.

The set of methodologies that constitute statistics includes mathematical, computational, and graphic methods that may be applied to a wide variety of data types including traditional numerical data, categorical data, image data, and even text data.'

P405 p.31

'The first U.S. Census was taken under the authority of Secretary of State Thomas Jefferson in 1790. U. S. Marshals on horseback ~~took the Census and they~~ counted 3.9 million people. By 1810, the U. S. Census was expanded to obtain information on manufacturing including the amount and value of products. By 1839, the American Statistical Society was formed to be renamed shortly the American Statistical Association because of an unfortunate acronym. In England, William Farr (1807-1883), an early medical statistician, was the compiler of abstracts in the office of the Registrar General. Using data that he compiled along with methods earlier attributed to John Snow, he identified the source of the 1866 cholera epidemic as water from a particular well of the London Water Company. Meanwhile his contemporary, Ernst Engel (1821-1896) served from 1860 as Director of the Royal Prussian Statistical Bureau.

Back in the United States, Abraham Lincoln establishes the United States Department of Agriculture (USDA) in 1862. Lincoln refers to USDA as "the people's department." In 1863, the first crop report appears and the USDA Division of Statistics is established. U. S. Census Bureau employee Herman Hollerith invented tabulating card machines, which were first used in the 1890 census, which counted nearly 63 million people. In 1913, the U. S. Department of Labor is established along with the Bureau of Labor Statistics.

A major development took place in Europe in 1953 with the development of the European Statistical System'

P405 p.32

'(EUROSTAT), which, for'

'the first time, integrated statistics across all of Western Europe. In short, the roots of statistics as a *state science* continues to stimulate and motivate statisticians with continuing advances in survey research and sampling theory associated with survey research.

P405 p.18

'Generally for statisticians, the set of methodologies that comprise statistics include mathematical, computational, and graphical methods and may be applied to a wide variety of types of data including traditional numerical data, categorical data, image data, and even text data.'

[48] In this case, P403 seems extracted from P405 by removal of the European events, but leaving the redundant "In the United States."

P403 p.136 (cont)[49]

'**Premodern Period**.[50] One of the most interesting early examples of the application of statistical methods before 1900 is the so-called Trial of the Pyx, a procedure developed in England beginning in the twelfth century to test newly minted coins for adherence to a quality standard. Uncertainty in the production process meant that some measure of error, or variation, had to be allowed for.
The London (later Royal) Mint selected a sample of each day's coins to be reserved in a box called the Pyx for later trial. The trial allowed for

a tolerance in the weight of a single coin and, by linear extrapolation in the aggregate weight of the contents of the Pyx. In this way

the integrity of the entire coinage could be judged based on one sample.

The roots of modern statistical methodology can be traced to the mid-seventeenth century.

John Graunt's (1620-1674) *Natural and Political Observations upon Bills of Mortality* published in 1662, used spatial data and map layouts to make inferences about sex ratios and disease types based on death records.

Toward the end of the Pre-modern period, **Sir Francis Galton** (1822-1911), in his study of heredity, developed the concept of regression toward the mean, described as early as the 1870s, and in 1888 he established the concept of correlation. In 1889 he published *Natural Inheritance,* in which he formally described the notions of regression and correlation.

The Classical Period.
The Classical Period (1900-1985) is characterized by a shift from descriptive methods to an increasingly mathematical formulation of methodologies. Computation was a tedious procedure and data collection a relatively costly process. Thus in the classical period there was considerable emphasis on optimality so that data were used efficiently, and on mathematical simplicity so that computation could be done rapidly. Hallmarks of theory developed in this era include small data sets, manual computation and strong and often unverifiable assumptions.'

P405 p.18

'In the Pre-modern Period, one of the most interesting early examples of the recognition of variability is the so-called Trial of the Pyx. The Trial of the Pyx is a procedure for maintaining the integrity of newly minted coins in the United Kingdom (England). From shortly after the Norman Conquest (1066) in a procedure that has been essentially unchanged since 1282,
the London (later Royal) Mint selects a sample of each day's coins that are reserved in a box called the Pyx.
The earliest agreements between the mint and the monarchy stated that a certain tolerance would be allowed in the weight of a single coin and by linear extrapolation in the aggregate weight of the contents of the Pyx. Thus, earlier than 1100, there was a formalized methodology for allowance of uncertainty and a method by which the integrity of the entire coinage could be judged based on a sample in the presence of uncertainty in the production process.

The roots of modern statistical methodology can be traced to the mid-seventeenth century. The earliest inferences are to a large extent based on graphical methods that are later echoed in what is labeled above as the Future Period.
John Graunt's (1620-1674) *Natural and Political Observations upon Bills of Mortality* published in 1662 gathered and used spatial data and map layouts to make inferences about sex ratios and disease types based on the bills of mortality.'

P405 p.19

'Towards the end of the Pre-modern period, Sir Francis Galton (1822-1911), cousin to Charles Darwin, developed the concept of regression toward the mean, described as early as the 1870s, and in 1888 he established the concept of correlation. In 1889, he published *Natural Inheritance,* in which he formally described the notions of regression and correlation.

The Classical Period
The Classical Period (1900-1985) is characterized by a shift from descriptive methods to an increasingly mathematical formulation of methodologies. ~~It must be remembered~~ that computation was a tedious procedure and data collection a relatively costly process. For this reason, in the classical period there was considerable emphasis on optimality so that data were used efficiently, and on mathematical simplicity so that computation could be done rapidly. Hallmarks of theory developed in this era include small data sets, manual computation, and strong and often unverifiable assumptions.'

P405 pp.19-24 discussed Karl Pearson, William S. Gossett, Sir Ronald Fisher, Egon Pearson, Jerzy Neyman, Andrei Nikolaevich Kolmogorov, Prasanta Chandra Mahalanobis, Harold Hotelling, Carl Harald Cramér, Calyampudi Radhakrishnan Rao, Samuel Wilks, i.e., major statisticians, none of whom were mentioned in P403.
The only statisticians named were Graunt and Galton for the 19th century and Wegman for the 20th.

[49] The red-bracketed section was almost certainly edited from the text at right, for which a plausible antecedent was Stigler's *The History of Statistics,* **§X.** The ideas remained, but several rounds of editing reduced the amount of identical text.
[50] It seems *odd* to spend several paragraphs on pre-1750 events and then cover almost nothing in the immediately-following "Classical Period."

P403 p.136 (cont)

'**Modern Period**. The Modern Period, from 1962, was marked by a major transition in thinking. Prior to 1962 in the Classical Period the focus was on the development of what is now called confirmatory analysis. Hypothesis testing, estimation, regression analysis, and variants of these were the major methodologies.

In contrast to confirmatory analysis in which a statistical model is assumed and inference is made on the parameters of that model, exploratory data analysis (EDA) is predicated on the fact that one does not necessarily know that model assumptions actually hold for data under investigation. Because the data may not conform to the assumptions of the confirmatory analysis, inferences made with invalid model assumptions are subject to (potentially gross) errors. The idea is to explore the data to verify that the model assumptions actually hold for the data in hand. This concept sparked a major revolution in the thought processes of statisticians.

Computers dramatically increased the statistician's ability to work with large quantities of data at greater levels of complexity and to analyze an interpret data faster and more efficiently.
The mid-1970s saw the emergence of integrated circuits and their use in primitive microcomputers,

but it was not until the IBM personal computer was introduced in 1981, and the Apple Macintosh was introduced in 1984, that serious computer power was in the hands of individual users. The introduction of personal computers and workstations in the 1980s dramatically increased access to computational resources, resulting an explosion of new statistical methods.'

P405 p.24-25

'**The Recent Past Period** The Recent Past Period (1962-2005) was marked by a major transition in thinking. Prior to 1962 in the Classical Period the focus was on the development of what is now called confirmatory analysis. Hypothesis testing, estimation, regression analysis, and variants of them were the major methodologies. ~~As mentioned earlier, these methods usually required strong and often unverifiable assumptions. John Tukey (1915-2000) represents a bridge between the Classical Period and the Recent Past Period. In the landmark 1962 paper of Tukey entitled, "The future of data analysis," and later in the 1977 book, *Exploratory Data Analysis*,[iii] Tukey sets forth a new paradigm for statistical analysis~~. In contrast to confirmatory analysis in which a statistical model is assumed and inference is made on the parameters of that model, exploratory data analysis (EDA) is predicated on the fact that one does not necessarily know that model assumptions actually hold for data under investigation. Because the data may not conform to the assumptions of the confirmatory analysis, inferences made with invalid model assumptions are subject to (potentially gross) errors. The idea ~~then~~ is to explore the data to verify that the model assumptions actually hold for the data in hand. This concept sparked a major revolution in the thought processes of statisticians and stimulated an outpouring of new methods.'

P405 pp.25-26 covered John Tukey, but also mentioned his many colleagues at Bell Labs and elsewhere. Tukey got no mention in P403.

P405 p.26-27

'The Future Period

' The introduction of personal computers and workstations circa 1981 sparked the beginnings of the Future Period (1981 onwards). In some ways it seems strange to date the Future from 1981, but the access to computational resources became so dramatically different, that literally an 'explosion of new methods resulted. … The placement of computer power in the hands of the end user made an enormous change in productivity. It should be noted that in the EDA table above the 1980-1984 and 1985-1989 period saw an explosion in papers in these two periods directly attributable to the introduction of personal computing.
The mid-1970s saw the emergence of integrated circuits and their use in primitive microcomputers. Indeed the first widely distributed microprocessor-based computer, Altair 8800, was announced in December of 1974. By July of 1976, the Apple I computer is introduced. Clearly a revolution was afoot, but it was not until the IBM personal computer, the SUN and Apollo Workstations in 1981 and the Apple Macintosh in 1984, that serious computer power was in the hands of individual users.'

P405 pp.27-30 then spent a page on Wegman, followed by paragraphs on Bradley Efron, Jerome Friedman, David W. Scott, with mentions of others.

P403 p.136 (cont)

'The contemporary period has also witnessed a clear change in research emphases. The post-*Sputnik* era (1957-1979) saw relatively lavish funding of basic research in statistics and an increasing emphasis on the development of methodology. However, the post-1981 era saw a significant shift in emphasis to applications. Computers allowed for new data structures, or methods of organizing information, many of which did not follow traditional statistical models. **Edward J. Wegman** called for the statistical profession to become more data-centric rather than methodology-centric: that is, to take on challenges of the new data structures even though they did not fit conveniently within the framework of existing statistical models. Some emerging data structures and future directions for the profession include streaming data, image data, text data, and data available in the form of random graphs. No longer is basic research money easily available for research in statistical methodology alone. Increased emphasis on real problems cannot help but be good for the discipline, because

P403 p.137 (cont)

'virtually every significant advance in statistics has been motivated by addressing some real problem.

At the beginning of the twenty-first century, forensics continued to emerge as an important new focus of statistics. Statistical methods have been used to discredit to a large extent the use of polygraphs for lie detection and exam results are rarely admisable as evidence in U.S. courts. Similarly, the National Research Council of the National Academies studied, using statistical methods, the use of bullet lead analysis by the Federal Bureau of Investigation resulting i increased legal challenges to this type of evidence. Other forensic science evidence likely to come under statistical and other technical scrutiny in the future includes what is now called friction ridge evidence and blood alcohol concentration evidence. Though DNA evidence has been shown to be valid from a statistical perspective, the statistical certainty of these other forms of forensic evidence is far less clear and is likely to lead to additional significant

adjustment in legal procedures and less aggressive pursuit of convictions based on these methods.'

P403 p.137 (cont)

'Assessment. The rise of statistics as a research field and as a vital component of statecraft is an important aspect of modern history, first in the West, then more globally. The importance of the discipline for key institutions like insurance companies, where actuarial work began to expand from the late nineteenth century onward, ensured its development. Arguably, popular training in statistics – as opposed to more conventional mathematics-has lagged somewhat in many societies, creating gaps in interpreting the results of data, and in some instances, significant disagreements over calculations of risk.

P405 p.30 (cont)

'The Future Period[51] is clearly changing the research emphases. The post-Sputnik era (1957-1979) saw relatively lavish funding of basic research in statistics with only some lip service being paid to applications. This substantial funding of undirected basic research saw also increasing emphasis on the development of methodology. However, the post-1981 era saw a significant increase in emphasis on applications. The availability of computing also resulted in new and novel data structures, many of which did not follow traditional statistical models. Wegman (2000) called for the statistical profession to become more data centric rather than methodology centric, i.e. to take on challenges of the new data structure even though they did not fit conveniently within the framework of existing statistical models. Some emerging data structures and future directions for the profession include streaming data, image data, text data, and data available in the form of random graphs. No longer is basic research money easily available for research in statistical methodology alone. Increased emphasis on real problems cannot help but be a good feature for academic research because

virtually every significant advance has been motivated by addressing some real problem.'

P405 p.32

An interesting new direction has been emerging with respect to forensics in the courtroom. Statistical methods have been used to discredit to a large extent the use of polygraph for lie detection and such testimony is no longer allowed (National Research Council, 2003). Similarly, the National Research Council of the National Academies (2004) has considered bullet lead analysis used by the Federal Bureau of Investigation using statistical methods and has increased legal challenges to this type of evidence. Other *forensic science* evidence likely to come under statistical and other technical scrutiny in the future include what is now called friction ridge evidence and blood alcohol concentration evidence. While DNA evidence has been vetted from a statistical perspective, the statistical certainty of these other forms of forensic evidence is far less clear and is likely to lead to additional significant

P405 p.33

'adjustment in legal procedures and less aggressive pursuit of convictions based on these methods.'

==========

P403 cited 5 sources[52], of which 3 were found in P405: NRC(2004), NRC(2003) and Wegman(2000). The following were new:

'David, H.A., and A. W. F. Edwards. Annotated Readings in the History of Statistics. New York: Springer, 2001.'

'Schweber, Libby. Disciplining Statistics: Demograph and Vital Statistics in France and England, 1830-1885. Durham, N.C. Duke University Press, 2006.'

[51] Historians might be *uncomfortable* labeling a time as a "Future Period."

[52] Stigler was not cited.

Z. GMU Proposal to ARO – 02/16/09

'Mathematical and Statistical Foundations of Networks'

Z.1 Proposal, $529K direct, labeled WEG2009

Federal agencies generally make proposals public if they are accepted, but (properly) not if they are rejected. Dan Vergano's October 2010 FOIA request elicited many files from Wegman related to an apparent proposal by him to ARO, with several folders:

ARO_Proposal

This working folder had many files, including a copy of Sharabati's dissertation, "MULTI-MODE AND EVOLUTIONARY NETWORKS" (2008), as "Dissertation.pdf," [SHA2008], as in [MAS2012c §4.4].

ARO Proposal

This folder's files were created 02/16/09, with the various pieces of a detailed proposal, including Table of Contents,[53] abstract, [54] biography[55], budget,[56] bibliography[57] and the 27-page technical proposal itself.[58]
Wegman listed 8 publications. Of the "5 publications most closely related," none were peer-reviewed, and he led only one, on alcoholism:
***P163* w** (2004), Martinez, A.R., **Wegman,** and Martinez, W.L.
P170w (2005), Solka, Bryant, Avory C., and **Wegman**
P178 Sw (2007), Said, **Wegman,** Sharabati, and Rigsby.
P179 Sw (2007, retracted), Said, **Wegman,** Sharabati, and Rigsby.
P192 Ws (2008). **Wegman** and Said.

The 5-year budget specified $322K to Wegman and $134K to Said, each 0.33 FTE, plus $37K for domestic travel, and other costs for total $529K. That did not include the usual ~48% added by GMU for indirect costs.

When funders evaluate proposals, they consider past performance, and *it is possible that the results and late reports of 0447 and 0059 were not plusses. However, from emails, ARO was generally encouraging and helpful to Wegman, perhaps given the long association.*
Some chronology may be worth reviewing:

11/01/04	0447 start
12/15/06	0059 start (of last ARO grant to Wegman)
12/15/07	0059 completion
04/30/08	0447 completion, after 6-month extension
12/10/08	0447 final report, 224 days after completion
12/15/08	Wegman: "I will be drafting something in the next week"
02/16/09	This proposal, [WEG2009]
03/08/09	0059 final report, 450 days after completion, cursory **§R.4**
05/07/09	negative feedback from ARO
05/29/09	5876 completion (Said)

§Z.2 gives some of the email history that shows the proposal was written in a month or two. He wrote that lack of support over previous summer caused him to have to take a second mortgage.
The emails also show some review comments, which were *as a group were strongly negative.*[59] One criticized a flaw in basic graph theory. Another said that many of the ideas had been explored before and that this proposal's lack of references to earlier work was not encouraging.[60]
In any case, the proposal was rejected, likely fortunate for Wegman, given the NSF Career Writing Workshop at GMU, 2009, p.66:[61]

> 'Plagiarism – material copied without citation and quotation – if you copy it, cite it and off-set it: **if you accept an award based on a proposal that includes plagiarism, you may have committed a felony**.'

[53] www.documentcloud.org/documents/524443-toc-network-science-proposal.html
[54] www.documentcloud.org/documents/524438-abstract-network-science-proposal.html
[55] www.documentcloud.org/documents/524440-biskchf-network-science-proposal-1.html
[56] www.documentcloud.org/documents/524435-masonbudget-aro-wegman-federal.html
[57] www.documentcloud.org/documents/524439-bibgraphyf-network-science-proposalpdf.html
[58] www.documentcloud.org/documents/524550-descriptf-network-science-proposal.html

[59] They are polite, but anyone familiar with reviews would understand the strength of the negatives, especially for a well-known, experienced researcher, not a proposal neophyte.
[60] This was an example of a pattern seen elsewhere, as Wegman and students *seemed to jump into unfamiliar areas* without studying them deeply first.
[61] grants.soe.ucsc.edu/sites/default/files/2%20George.ppt

Z.2 Emails

[WEG2010a][62] contains some back history on this proposal.
Some messages are excerpted here in chronological order.

10/01/08 p.17 Wegman to ARL

'Great to hear from you. Yes I am sure he would be willing to do so. He is my Ph.D. student and will be defending his dissertation shortly. The final defnse *(sic)* will be in about four weeks. **I am still hoping we can get into a new contract arrangement with ARL. This past summer I had no research support, which is very hard for me.**'

12/10/08 pp.21-22 ARO to Wegman on 0447

'Your Final Report has been received. …
U.S. Army Research Office …
DATES COVERED: 1-Nov-2004 to 30-Apr-2008 …
PROPOSAL TITLE: Analytic and Graphical Methods for Streaming Data with Applications to netcentric Warfare'

12/15/08 p.20 Wegman to ARO

'**Now that I have finally turned in my final report**,[63] I'd like a little advice. I want to submit a new proposal. I believe the last effort was very fruitful and I have some good ideas. I want to work on the mathematical foundations of network science and have several ideas on how to detect missing nodes and edges and also how to deal mathematically with dynamically expanding networks. I hope this is of interest to ARO. **Last summer, I went without research support and wound up having to take a second mortgage on my home in order to make ends meet.** I have been very supportive of the Arny nission *(sic)* over the years and ARO has returned the favor by being supportive of me. Without extracting a commitment from you ahead of time, I wonder if you might give me some guidance on possible funding levels. **I will be drafting something in the next week**, but I'd like to have some sense of what might be feasible. **For myself I would like summer support and perhaps a little released time during the year**. **Would it be possible to include some support for Dr. Yasmin Said who has been working closely with me over the last 2 ½ years?**'

02/19/09 p.23 GMU to Wegman

'Your application has been submitted to U.S Army Research Office on Grants.gov. Attached is the submission confirmation.'

05/07/09 pp.27-28 ARO to Wegman (entirely quoted, 4:1 *negative*)

'Prof Wegman,
It was good to have a conversation about the proposed project Mathematical and Statistical Foundations of Networks. I did contact the Army reviewer yesterday. She assures me she will work on the review.
Here are a comments from reviewers that I find helpful. Please do not take these comments as negative. They are intended to help improve the proposal.

"Very important topic, prominent investigators, however, **the description of the ideas is rather vague and still needs to be developed into a sound research plan. It is not clear how addressing the proposed research tasks will advance the existing theory behind network science."** *negative*

"Some of the proposed ideas … For example, the authors propose a method to estimate the probability of missing edges. Then they propose to use this method for estimating the probability of missing nodes for G by estimating the probabilities for corresponding edges in the line graph representation of G. **The problem is that, unless G is claw-free,[64] there may not be a one-to-one correspondence between vertices in G and edges in its line graph** representation (even though there is a one-to-one correspondence between edges in G and vertices in its line graph, this is not what one needs in this case)." *negative (claw-free comment applies to section 2.2.3 of the proposal)*

"The proposal outlines some potential new capabilities that can arise from the proposed study. **Some of those are very interesting**, like task **2** focusing on conversion of multimode non-binary adjacency tensors and matrices into lower degree networks or evolutionary algorithms for optimizing network assessment metrics. **Other tasks,** like **3** (evolving social networks), **4** (missing edges), 7 (connection between text mining and social networks), or **8** (limiting behavior of agent-based systems) **have been already studied and lack of references to the relevant work[65] makes this reviewer doubtful of the likelihood that the proposed research will lead to new capabilities in this tasks."** *negative*

"**With better description of the methods to be applied and perhaps some initial results showing the promise of these methods, as well as with clearly defined expected results, the proposal will be much stronger**." *negative*

"The PI is the distinguished scientist with strong past experience and publication record of papers relevant to the proposed study."' *positive*

[62] www.documentcloud.org/documents/527437-relevant-emails-redact-addresses.html
[63] That was the **0447** final report. The **0059** final report was **03/18/09,** ~year late.
[64] en.wikipedia.org/wiki/Claw-free_graph Well-known graph theory.
[65] [MAS2010a §W.5] noted this issue for Wegman group's SNA.

Z.3 Pages of proposal, side-by-side comparisons

1-2 Most of this text is quoted (properly) from an NRC report, but of course provides no information on the proposal itself.

2-9 Most is taken from Walid Sharabati's dissertation [SHA2008], co-supervised by Wegman and Said. About 2 pages were plagiarized from Wikipedia, [DEN2005] and especially [WAS1994].

10-15 Adds Iraq War, evolving networks, but ~half is from [SHA2008].

15-17 Most is almost identical to text and definitions from Hadi Rezazad's dissertation [REZ2009], Spring 2009, a few months later. However, this text must have been done much earlier.

18-21 No/few antecedents were found.

22- Mostly unknown, some passages from Wikipedia.

24-25 "Ising Models" is nearly identical to English in Chinese Wiki. Both came *likely* from Binder or Brush below, which Wegman cited as "See Brush (1967) and Binder (2001)," but with no hint that the text might have been copied from them.

26-27 Summary of tasks, no antecedents found.

Total About half the text appears to have been copied without proper attribution,[66] leaving half as possible new content. *Some may ignore the use of Wegman-supervised PhD dissertations,* but ~4 pages were copied from others besides. The dissertations were not mentioned.

The bibliography[67] seems a *strange* mix of textbooks and sometimes-obscure references, with few relevant, recent papers to show field familiarity:

'Binder, K. (2001) "Ising model," in Encyclopaedia of Mathematics (Hazewinkel, Michiel, ed.), Kluwer Academic Publishers. *Vaguely cited in the Summary,*

Brush, Stephen G. (1967) "History of the Lenz-Ising model," Reviews of Modern Physics (American Physical Society) 39, 883–893. doi: 10.1103/RevModPhys.39.883 *Vaguely cited in the Summary,*

Committee on Network Science for Future Army Applications (2005) Network Science, Washington, DC: The National Academies Press. *Properly cited.*

de Nooy, W., Batageli, V., and Mrvar, A. (2004) Exploratory Social Network Analysis with Pajek, Cambridge, UK: Cambridge University Press. *[DEN2005] cited "for more details refer to de Nooy et al. (2004)"*

Epstein, Joshua and Axtel, Robert (1996) Growing Artificial Societies: Social Science from the Bottom UP (Complex Adaptive Systems), Washington, DC: The Brookings Institution. *Cited once, vaguely.*

Marchette, David J. and Priebe, Carey E. (2008) "Predicting unobserved links in incompletely observed networks," Computational Statistics and Data Analysis, 52(3), 1373-1386. *Cited once.*

Martinez, Angel (2002) A Framework for the Representation of Semantics, Ph.D. Dissertation, School of Computational Sciences, George Mason University, Edward J. Wegman, Dissertation Director. *Cited 2 times.*

Martinez, Wendy, Martinez, Angel, and Wegman, Edward (2008) "Classification and clustering using weighted text proximity matrices," Computing Science and Statistics, 36, 600-611. *Interface proceedings*, *Cited once.*

North, Michael and Macal, Charles (2007) Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation, New York: Oxford University Press. *Cited once, vaguely.*

Said, Yasmin (2009) Intervention to Prevention: A Policy Tool for Alcohol Studies, Saarbrücken, Germany: VDM Verlag. Dissertation, *See **§S.8*** 17.

Solka, Jeffrey L., Bryant, Avory C., and Wegman, Edward J. (2005) "Text data mining with minimal spanning trees," Handbook of Statistics: Data Mining and Data Visualization, (Rao, C. R., Wegman, E. J. and Solka, J. L., eds.), 133-170. Amsterdam: Elsevier/North Holland. P170w *Cited once.*

van Rooij, A. and Wilf, H. (1965) "The interchange graph of a finite graph," Acta Math. Acad. Sci. Hungar. 16, 263-269.

Wasserman, Stanley and Faust, Katherine (1994) Social Network Analysis: Methods and Applications, Cambridge, UK: Cambridge University Press.' *[WAS1994] This is actually cited 3 times, but with no quotations. Indeed, much text came from there but in 2011, Wegman was claiming he'd thought it had been original text by his student Reeves. The sequence was as follows:*

[66] If text is copied, it must be quoted and cited, not vaguely mentioned nearby.
[67] www.documentcloud.org/documents/524439-bibgraphyf-network-science-proposalpdf.html

[WEG2009] ← [SHA2008] ← Sharabati ← Wegman ← Reeves ← {sources} Wegman wrote to Elsevier about the origin of the SNA text[68] used in the WR and [SAI2008]. [MAS2011a, p.6-7]:

> 'I asked her *(Reeves)* to write up a short description we could include in our summary. She provided that within a few days, **which I of course took to be her original work.** Neither Yasmin, Walid Sharabati, John Rigsby nor I did know that she had basically copied and pasted this into her MS Word file. We included her boilerplate in our Congressional testimony and acknowledged Denise's contribution in that testimony. …'[69]
> '**thinking that the page and ½ Denise had given me was original work** that had not been formally published, I gave it as reading material to Walid as background material along with a number of other references. **Walid included it as background material in his dissertation with only minor amendments**.'

Some of the SNA text appeared in WR, [SAI2008], [SHA2008] and [REZ2009]. [WEG2009] was the 5th known re-use of Reeves' text, **§G.** Wegman seems to have re-used Rezazad's work later. The reader can compare Rezazad[70] last modified 09/30/09, on material from his dissertation, with him as contact and Wegman, Rezazad, Shores[71] dated 10/22/09 with Wegman as contact. It edited "we" ← "I" on a few pages, and added pp.34-40, i.e., graph theory that *seems a bit extraneous*.

Wegman was entirely responsible for this proposal to the ARO. No grad students were ack'd or involved. Hence, there can be no doubt as to the authorship.

[VER2011] quoted Milton Johns and Wegman:

> '"**Neither Dr. Wegman nor Dr. Said has ever engaged in plagiarism**," says their attorney, Milton Johns, by e-mail. In a March 16 e-mail to the journal, Wegman blamed a student who "had basically copied and pasted" from others' work into the 2006 congressional report, and said the text was lifted without acknowledgment and used in the journal study. "**We would never knowingly publish plagiarized material**" wrote Wegman, a former *CSDA* journal editor.'

§Z.3 gives evidence of plagiarism, in the side-by-side comparison style used elsewhere by DC and [MAS2010a], adapted as needed for 3-way:

- Cyan for identical, in-order text between [WEG2009] and antecedents
- Yellow for trivial edits [WEG2009] ← [SHA2008]
- Yellow for trivial edits [SHA2008] ← antecedents, shown only in antecedents, unlike in two-way comparisons.

Page numbers use the document's own numbers, not those of the PDF. When mosaic plagiarism[72] is suspected in document X, if closely-matching text in Y can be found for a section of X then either:

- Y is the antecedent, either directly or indirectly OR
- X and Y both have another hidden antecedent OR
- Possibly X is original and was a source for Y. This always needs checking, but publications dates and/or authors' expertise help.

If no matching text can be found:

- X is original or well-paraphrased from Y OR
- A hidden antecedent source has not yet been identified.

[68] Deep Climate, —A comparison of Said, Wegman, et al and Unattributed Sources, 09/08/10. deepclimate.files.wordpress.com/2010/09/said-et-al-social-networks-2.pdf

[69] The SNA introduction in the WR was ~5.5 pages, *likely* more than the work of the 2nd author, Scott, but she was only vaguely Ack'd, not labeled an author.

[70] "ACAS 2009\GMU-Presentation-092509.ppt" www.documentcloud.org/documents/550209-gmu-presentation-092509.html

[71] "ACAS_2009\ACAS_Wegman_Rezazad\Shores.ppt" www.documentcloud.org/documents/550210-acas-wegman-rezazad-shores.html

[72] isites.harvard.edu/icb/icb.do?keyword=k70847&pageid=icb.page342054#a_icb_pagecontent732741_mosaic for example.

[WEG2009, pp.1-2]
Most of this quotes excerpts from a 2005 NRC report on Network Science, acceptably referenced, but not original.

[WEG2009, p.2[
2 SOCIAL NETWORKS
Social Network Analysis (SNA) or Network Theory is becoming an important tool used to analyze, model, and simulate the behavior of groups of people or entities both on the global level (how two or more groups interact with other group(s)) and on the local level (how individuals interact with each other within the same network.) In the past two decades, SNA has been used to analyze relations and ties among individuals of the same network and similarities between different networks to obtain a better understanding on how societies interact.

The basic mathematical structure for visualizing the social network is a *graph*. A graph is a pair {V;E} where V is a set of nodes or vertices and E is a set of edges or links
Social network analysis

has emerged as a key technique and a topic of study in modern sociology, anthropology, social psychology and organizational theory.
The shape of the social network helps determine a network's usefulness to its individuals. Smaller, tighter networks can be less useful to their members than networks with lots of loose connections (weak ties) to individuals outside the main network. More "open" networks,

[WEG2009, p.3[
with many weak ties and social connections, are more likely to introduce new ideas and opportunities to their members than closed networks with many redundant ties.

[SHA2008, p.1]

Social Network Analysis (SNA) or Network Theory is becoming important tools used to analyze, model, and simulate the behavior of groups of people or entities both on the global level (how two or more groups interact with other group(s)) and on the local level (how individuals interact with each other within the same network.) In the past two decades, SNA has been used to analyze relations and ties among individuals of the same network and similarities between different networks to obtain a better understanding on how societies interact.

[SHA2008, p.2]
The basic mathematical structure for visualizing the social network is a graph. A graph is a pair V;E where V is a set of nodes or vertices and E is a set of edges or links.
Social network analysis ~~(also called network theory)~~

has emerged as a key technique and a topic of study in modern sociology, anthropology, social psychology and organizational theory.
The shape of the social network helps determine a network's usefulness to its individuals. Smaller, tighter networks can be less useful to their members than networks with lots of loose connections (weak ties) to individuals outside the main network. More “open” networks,

with many weak ties and social connections, are more likely to introduce new ideas and opportunities to their members than closed networks with many redundant ties.

Original antecedents, most from [WAS1994], some from Wikipedia or [DEN2005].

Wikipedia – Social Networks – 01/02/06[73]
[Head section 1]

Social network analysis (also ~~sometimes~~ called network theory)
has emerged as a key technique in modern sociology, anthropology, Social Psychology and organizational studies, ~~as well as a popular topic of speculation and study.~~
The shape of the social network helps determine a network's usefulness to its individuals. Smaller, tighter networks can be less useful to their members than networks with lots of loose connections (weak ties) to individuals outside the main network. More "open" networks,

with many weak ties and social connections, are more likely to introduce new ideas and opportunities to their members than closed networks with many redundant ties.

[73] en.wikipedia.org/w/index.php?title=Social_network&oldid=33590649 called [WIK2006a] elsewhere.

[WEG2009, p.3 continued]

Social network analysis is concerned with understanding the linkages among social entities and the implications of these linkages. The social entities are referred to as actors that are represented by the vertices of the graph.

Most social network applications consider a collection of actors that are all of the same type. These are known as one-mode networks.

Social ties link actors to one another.
The range and type of social ties can be quite extensive.

Linkages are represented by edges of the graph. Examples of linkages include the
evaluation of one person by another (such as expressed friendship, liking, respect),
transfer of material resources (such as business transactions, lending or borrowing things),
association or affiliation (such as jointly attending the same social event or

belonging to the same social club),
behavioral interaction (talking together, sending messages),
movement between places or statues ★
(migration, social or physical mobility),
physical connection (a road, river,
bridge connecting two points),
formal relations such as authority, and
biological relationships such as kinship or descent.

The tie is an inherent property of a pair.

Many kinds of network analysis are concerned with understanding ties among pairs and are based on the *dyad* as the unit of analysis.

The "statues" typo (or bad original OCR by Reeves) got carried through all re-uses of [WAS1994] except [SAI2008].

[SHA2008, p.2]

Social network analysis is concerned with understanding the linkages among social entities and the implications of these linkages. The social entities are referred to as actors that are represented by the vertices of the graph.

Most social network applications consider a collection of actors that are all of the same type. These are known as one-mode networks

Social ties link actors to one another.
The range and type of social ties can be quite extensive.
~~A tie establishes a linkage between a pair of actors.~~

Linkages are represented by edges of the graph. Examples of linkages include the
evaluation of one person by another (such as expressed friendship, liking, respect),
transfer of material resources (such as business transactions, lending or borrowing things),
association or affiliation (such as jointly attending the same social event or

[SHA2008, p.3]

belonging to the same social club),
behavioral interaction (talking together, sending messages),
movement between places or statues ★
(migration, social or physical mobility),
physical connection (a road, river,
bridge connecting two points),
formal relations such as authority and
biological relationships such as kinship or descent.
~~A linkage or relationship establishes a tie at the most basic level between a pair of actors.~~
The tie is an inherent property of the pair.

Many kinds of network analysis are concerned with understanding ties among pairs and are based on the dyad as the unit of analysis.

[WAS1994, p.17]

Actor.
social network analysis is concerned with understanding the linkages among social entities and the implications of these linkages. The social entities are referred to as actors. ~~Our use of the term "actor" does not imply that these entities necessarily have volition or the ability to "act". Further~~
most social network applications focus on collections of actors that are all of the same type …. We call such collections one-mode networks…

[WAS1994, p.18]

Relational tie. Actors are linked to one another by social ties. …the range and type of ties can be quite extensive. The defining feature of a tie ~~is that it~~ establishes a linkage between a pair of actors. Some of the more common examples of ties employed in network analysis are:

- Evaluation of one person by another (for example expressed friendship, liking, or respect)
- Transfers of material resources (for example business transactions, lending or borrowing things)
- Association or affiliation (for example jointly attending a social event, or belonging to the same social club)
- Behavioral interaction (talking together, sending messages)
- Movement between places or statuses ★ (migration, social or physical mobility)
- Physical connection {a road. river, ~~or~~ bridge connecting two points}
- Formal relations (for example authority)
- Biological relationship (kinship or descent)

Dyad. At the most basic level, a linkage or relationship establishes a tie between two actors.
The tie is inherently a property of the pair ~~and therefore is not thought of as pertaining simply to an individual actor.~~

Many kinds of network analysis are concerned with understanding ties among pairs. All of these approaches take the dyad as the unit of analysis

[WEG2009, p.3 cont]

A social network consists of a finite set or sets of actors and the relation or relations defined on them. The presence of relational information is a significant feature of a social network.

A partition of a network is a classification or clustering of the vertices in the network so that each vertex is assigned to exactly one class or cluster; that is,

partitions divide the vertices of a network into a number of mutually exclusive subsets. Partitions may specify some property that depends on attributes of the vertices.
Partitions are also sometimes called blocks or block models. These are essentially a way to cluster actors together in groups that behave in a similar way.

In a network setting, actors or entities have several attributes to identify their role, behavior, background, and/or assets; some of which are unique to that actor and some are common among other actors. These attributes are the node's properties such as gender, age, political affiliation, ethnicity, race, nationality, religion, spoken languages, scientific field, income, education level, job class, and geographic location.

2.1 PRELIMINARIES

[WEG2009, p.4]
Networks can be treated as directed graphs.

There are three basic representations of a network: the planar graph visualization, the adjacency matrix, and the sparse-graph representation.
There are several algorithms to study interactions within the network include centrality measures (node degree and closeness), network partitioning (cliques and clique overlapping), network connectivity (cut-points and bridges), structural *equivalence, structural holes, brokerage roles, and block-modeling.*

[SHA2008, p.3 cont]

A social network consists of a finite set or sets of actors and the relation or relations defined on them. The presence of relational information is a significant feature of a social network.
"significant" ← *"critical&defining" not same.*

A partition of a network is a classification or clustering of the vertices in the network so that each vertex is assigned to exactly one class or cluster. Partitions may specify some property that depends on attributes of the vertices.
Partitions divide the vertices of a network into a number of mutually exclusive subsets. ~~That is, a partition splits a network into parts.~~
Partitions are also sometimes called blocks or block models. These are essentially a way to cluster actors together in groups that behave in a similar way.

In a network setting, actors or entities have several attributes to identify their role, behavior, background, and/or assets; some of which are unique to that actor and some are common among other actors. These attributes are the nodes' properties such as gender, age, political affiliation, ethnicity, race, nationality, religion, spoken languages, scientific field, income, education level, job class, and geographic location.

[SHA2008, p.4]
1.2 Preliminaries

Networks can be treated as directed graphs in which actors (individuals) are represented by vertices (nodes) while interactions between actors are represented by edges (ties), which may have weights.
There are three basic representations of a network {the planar graph visualization, the adjacency matrix, and the sparse-graph representation.
There are several algorithms to study interactions within the network include centrality measures (node degree and closeness), network partitioning (cliques and clique overlapping), network connectivity (cut-points and bridges), structural equivalence, structural holes, brokerage roles and block-modeling, ~~which will all be defined shortly.~~

[WAS1994, p.20]

Social Network. Having defined actor, group, and relation we can now give a more explicit definition of social network
A social network consists of a finite set or sets of actors and the relation or relations defined on them. The presence of relational information is a critical and defining feature of a social network …

[DEN2005, p. 31]
A partition of a network is a classification or clustering of the vertices in the network so that each vertex is assigned to exactly one class or cluster.
[DEN2005, p. 36]
Partitions divide the vertices of a network into a number of mutually exclusive subsets. In other words, a partition splits a network into parts.

From just these short passages alone, it would not be obvious that [DEN2005] was the original antecedent.

However, [MAS2010a §W.2.3] showed these as a few of many [DEN2005] passages used in the WR, of which only a small subset got re-used..

[WEG2009, p.4 cont]

Definition 2.1. A graph, G, is a collection of vertices V and edges E; G = {V, E} …
Definition 2.2**.** An adjacency matrix, E, associated with a graph, G
(slightly simplified notation)

A key insight of these definitions is that there is a fundamental duality between the graph and its adjacency matrix. That is, if one is given the adjacency matrix, one is able to construct the graph, and similarly, if one is given the graph, one can construct the adjacency matrix. The computationally-oriented social scientists tend to focus on the graph because it is a descriptive way of visualizing the social network. The adjacency matrix is used, but in SNA, the properties of the network are rarely explored in the context of a deeper mathematical analysis of the matrix representation. Because much is known about linear spaces, it is the theme of this proposal that we can understand much about networks in terms of this matrix representation. There are a number of metrics that describe quantitative aspects of a social network. In particular, there is much discussion of dyads, triads and cliques in social network analysis.

2.1.1 CENTRALITY MEASURES
There are three main centrality measures defined in Wasserman and Faust (1994); namely, degree centrality, closeness centrality, and betweenness centrality.

Degree of a vertex is the number of edges that connect it to other nodes. Degree can be interpreted as measure of power or importance of a node, or measure of workload. The

actor with most ties is the most important actor in a network. It has been shown that in a simple random graph, degree centrality has the Poisson distribution. Nodes with high degree are likely to

[WEG2009, p.5]
be at the intuitive center. Deviations from a Poisson distribution suggest non-random processes, such processes form *scale-free networks*.

[SHA2008, p.4 cont]

Definition 1.2.1. A graph G, is a collection of vertices V and edges E; G = {V;E}, ….
Defnition 1.2.2. An adjacency matrix A associated with a graph G
(standard definitions of graphs, no problem for Sharabati)

No antecedent found.

1.2.1 Centrality Measures
There are three main centrality measures defined in [60]; namely, degree centrality, closeness centrality and betweenness centrality. ~~To serve the purposes of this research, I will define degree and closeness centrality measures only.~~
Degree of a vertex is the number of edges that connect it to other nodes, ~~see Figure 1.1~~. Degree can be interpreted as measure of power or importance of a node, or measure of workload. The
[SHA2008, p.5]
actor with most ties is the most important figure in a network. It has been shown that in a simple random graph, degree centrality has the Poisson distribution. Nodes with high degree are likely to

be at the intuitive center. Deviations from a Poisson distribution suggest non-random processes, such processes form "scale-free" networks

The reader might wonder if well-published SNA experts would agree with the opinions at left.
Google: "social network analysis" "adjacency matrix" OR
"social network analysis" "adjacency matrix" multimode OR
"social network analysis" "adjacency matrix" cuboid

It is curious that references to [WAS1994] and [DEN2005] appeared in [SHA2008], but no one seemed to worry that the original SNA text in the WR might have been plagiarized. In any case, Wegman had claimed they all thought this was original work of Reeves.

Wegman and some of his students seemed to have a habit of copying text that gave fairly-standard basic definitions, many of which were not even used in the remainder of their text. It seems that this was just intended to convey the impression of expertise, and add bulk, because the results were useful neither to the general public (as in the WR) nor to experts. The latter would likely write "we use the standard terminology of <source>, notation summarized below"

[WEG2009, p.5 cont]
Definition 2.3. Degree of a vertex …
Definition 2.4. Closeness; …
(slightly simplified notation, but standard)
Closeness centrality measure is based on the inverse of the distance of each actor to every other actor in the network.
Distance in this context is defined to be the number of steps a vertex v_i needs to reach a vertex v_j . If an actor is close to all other actors then this actor is considered important.

Definition 2.5. The geodesic is the length of the shortest path between any two vertices.

2.1.2 COHESIVE SUB-GROUPS: CLIQUES
Definition 2.6. A dyad is a pair of vertices and the edge connecting them.

Definition 2.7. A triad is a set of three vertices and the edges connecting them.

A triad is identified by a M-A-N number system of three digits and a letter; for more details refer to de Nooy et al. (2004)
One of the interesting features in a network that caught structural analysts' attention is secondary sub-structures such as network cohesion.
[WEG2009, p.6]
Researchers interested in cohesive subgroups gathered and studied sociometric data on affective ties

in order to identify "cliquish" subgroups (face-to-face group).

The clique is the foundational idea for studying and analyzing cohesive subgroups in social networks.

Definition 2.8. A clique in a graph is a maximal complete subgraph of three or more nodes, mutual dyads are not considered to be cliques (Wasserman and Faust, 1994).
It consists of a subset of nodes all of which are adjacent to each other, and there are no other nodes that are also adjacent to all of the members of the clique.
A clique is a very strict definition of cohesive subgroups.
Cliques are a subset of the network in

which the actors are more closely and intensely tied to one another than they are to other members of the network and if one actor disappears for any reason, the others are still directly connected to each other.

[SHA2008, p.5 cont]
Definition 1.2.3. Degree of a vertex …
Definition 1.2.4. Closeness; …

Closeness centrality measure is based on the inverse of the distance of each actor to every other actor in the network.
Distance in this context is defined to be the number of steps a vertex v_i needs to reach a vertex v_j . If an actor is close to all other actors then this actor is considered important.
[SHA2008, p.6]
Definition 1.2.5. The geodesic is the length of the shortest path between any two vertices.

1.2.2 Cohesive Sub-Groups: Cliques
Definition 1.2.6. A dyad is a pair of vertices and the edge connecting them.

Definition 1.2.7. A triad is a set of three vertices and the edges connecting them.

A triad is identified by a M-A-N number system of three digits and a letter; for more details refer to [14].
One of the interesting features in a network that caught structural analysts' attention is secondary sub-structures such as network cohesion.

Researchers interested in cohesive subgroups gathered and studied sociometric data on affective ties

in order to identify "cliquish" subgroups (face-to-face group).

The clique is the foundational idea for studying and analyzing cohesive subgroups in social networks.

Definition 1.2.8. A clique in a graph is a maximal complete subgraph of three or more nodes, mutual dyads ~~(2 nodes)~~ are not considered to be cliques [60].
It consists of a subset of nodes all of which are adjacent to each other, and there are no other nodes that are also adjacent to all of the members of the clique.
A clique is a very strict definition of cohesive subgroups.
Cliques are a subset of the network in
[SHA2008, p.7]
which the actors are more closely and intensely tied to one another than they are to other members of the network and if one actor disappears for any reason, the other two can still write/talk to each other.

Standard definitions.

[WAS1994, p.253]
Researchers interested in cohesive subgroups gathered and studied sociometric data on affective ties,
such as friendship or liking in small face-to-face groups,
in order to identify "cliquish" subgroups …
[WAS1994, p.254]
The clique is the foundational idea for studying and analyzing cohesive subgroups in social networks.

A clique in a graph is a maximal complete subgraph of three or more nodes … mutual dyads are not considered to be cliques.
It consists of a subset of nodes all of which are adjacent to each other, and there are no other nodes that are also adjacent to all of the members of the clique.
A clique is a very strict definition of cohesive subgroups.

[SHA2008] at least had a vague citation to [WAS1994], but still was plagiarized.

[WEG2009, p.6 cont]

2.1.3 BLOCKMODELS

Definition 2.9. A blockmodel is the process of identifying positions in the network.

A block is a section of the adjacency matrix consisting of a group of actors that are structurally equivalent. It consists of two things according to Wasserman and Faust (1994):

A partition of actors in the network into discrete subsets called positions.

For each pair of positions a statement of the presence or absence of a tie within or between the positions on each of the relations.

A blockmodel is thus a hypothesis about a multirelational network. It presents general features of the network, such as the ties between positions, rather than information about individual actors.

A blockmodel is a simplified representation of multirelational network that captures some of the general features of a network's structure. Specifically, positions in a blockmodel contain actors who are approximately structurally equivalent. Actors in the same position have identical or similar ties to and from all actors in other positions.

Thus, the blockmodel is stated at the level of the positions, not individual actors.

2.1.4 TWO-MODE AND MULTIMODE NETWORKS

Some social network relationships can be treated as a two-mode "bipartite" networks, or three-mode "tripartite" networks. As an example, consider the author-paper networks,

[WEG2009, p.7]

there are two types of vertices, one class of vertices represents authors, while the other class of vertices represents papers. There is one relationship type; "person A authored/coauthored paper P". It is understood that in this case the adjacency matrix is binary, i.e. the entries in the matrix are only 0s and 1s. (Later we shall see situations where the entries in the adjacency matrix can represent strength of relationship, e.g. measured as frequency counts or probabilities.)

One can perform matrix operations such as the product of matrices to obtain interesting results.

[SHA2008, p.8]

1.2.4 Blockmodel

Definition 1.2.10. A blockmodel is the process of identifying positions in the network.

A block is a section of the adjacency matrix "a group of people" structurally equivalent. It consists of two things according to Wasserman and Faust [60]:

●A partition of actors in the network into discrete subsets called positions.

●For each pair of positions a statement of the presence or absence of a tie within or between the positions on each of the relations.

A blockmodel is thus a hypothesis about a multirelational network. It presents general features of the network, such as the ties between positions, rather than information about individual actors.

A blockmodel is a simplified representation of multirelational network that captures some of the general features of a network's structure. Specifically, positions in a blockmodel contain actors who are approximately structurally equivalent. Actors in the same position have identical or similar ties to and from all actors in other positions.

Thus, the blockmodel is stated at the level of the positions, not individual actors.

[SHA2008, p.10]

1.3.1 Relational Networks

Some social network relationships can be treated as a two-mode "bipartite" networks, or three-mode "tripartite" networks. As an example, consider the author-paper networks,

there are two types of vertices, one class of vertices represents authors while the other represents papers. There is one relationship type; "person A authored/coauthored paper P". ~~This two-mode relational socio-network can be concluded from the PCANS model [33], [9]. The PCANS model is presented in Table 1.1:~~

I can perform matrix operations such as the product of matrices to obtain interesting results given that the two-mode matrix is binary.

[WAS1994, p.395]

A blockmodel consists of two things:

(i)A partition of actors in the network into discrete subsets called positions.

(ii)For each pair of positions a statement of the presence or absence of a tie within or between the positions on each of the relations.

A blockmodel is thus a hypothesis about a multirelational network. It presents general features of the network, such as the ties between positions, rather than information about individual actors.

[WAS1994, p.395]

A blockmodel is a simplified representation of multirelational network that captures some of the general features of a network's structure. Specifically, positions in a blockmodel contain actors who are approximately structurally equivalent. Actors in the same position have identical or similar ties to and from all actors in other positions.

For example, all actors in position B_k have similar ties to actors in positions B_l, B_m, and so on.

Thus, the blockmodel is stated at the level of the positions, not individual actors.

After this point, no obvious text was found from [WAS1994] or elsewhere, so presentation switches to 2-column.

[WEG2009, p.7 cont]

2.1.4.1 EXAMPLE

Consider a bipartite "coauthor-by-paper" social network. Let A be the adjacency matrix of size m x n representing the graph of the network, with m = number of coauthors, and n = number of papers. Then ,

$C_{m \times m} = A_{m \times n} \cdot A^T_{n \times m}$ = the one-mode coauthorship adjacency matrix, and

$P_{n \times n} = A^T_{n \times m} \cdot A_{m \times n}$ = the one-mode paper-by-paper adjacency matrix

where

$c_{ii} = \sum_{j=1}^{n} a_{ij}$ = the number of papers author i published

$p_{jj} = \sum_{i=1}^{m} a_{ij}$ = the number of coauthors that coauthored paper j, and

c_{ij} = the tie-strength between authors i and j.

Finally if $D_{m \times m} = C^2_{m \times m}$, then d_{ii} = the vertex degree of author i.

Suppose the coauthor-by-paper adjacency matrix A is given by

	paper1	paper2	paper3
coauthor1	1	0	1
coauthor2	0	1	1
coauthor3	1	1	0
coauthor4	1	0	0

[WEG2009, p.8]

$$\Rightarrow A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}_{4 \times 3}$$

There is a one-to-one correspondence between the adjacency matrix representation of a network and directed graphs.

[SHA2008, p.11]

Consider a bipartite "coauthor-by-paper" social network. Let A be the adjacency matrix of size m x n representing the graph of the network, with m = number of coauthors, and n = number of papers. Then,

$C_{m \times m} = A_{m \times n} \cdot A^T_{n \times m}$ = coauthorship proximity matrix, and

$P_{n \times n} = A^T_{n \times m} \cdot A_{m \times n}$ = paper-by-paper proximity matrix.

where,

$c_{ii} = \sum_{j=1}^{n} a_{ij}$ = number of papers author i published,

$p_{jj} = \sum_{i=1}^{m} a_{ij}$ = number of coauthors coauthored paper j, and

c_{ij} = tie-strength between coauthors i and j.

Finally, if $D_{m \times m} = C^2_{m \times m}$ then

d_{ii} = vertex degree of coauthor i.

[SHA2008, p.11]

~~Example:~~

Suppose the coauthor-by-paper adjacency matrix A is given by

[htpb]

	paper1	paper2	paper3
coauthor1	1	0	1
coauthor2	0	1	1
coauthor3	1	1	0
coauthor4	1	0	0

$$\Rightarrow A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}_{4 \times 3}$$

There is a one-to-one correspondence between the matrix representation of a ~~social~~ network and directed graphs.

[WEG2009, p.8]

2.1.4.1 EXAMPLE

....

....

....

....

....

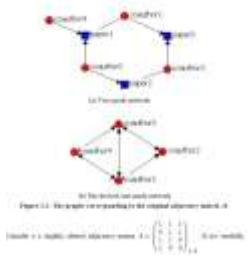
....

....

....

....

....



If we carefully examine the networks in Figures 2.1 and 2.2, we observe that these

different 2-mode networks in fact have the same 1-mode graphical network representation. This is due to the fact that when converting to a 1-mode network, some network features are lost; much.

[WEG2009, p.9]

the same effect when one projects from 3-D to 2-D. This is an example of how the one-mode network does not provide sufficient information on how peer-ties are formed. As a result, the analysis of two-mode networks, indeed, multi-mode networks and one-mode networks should be performed concurrently.

The blockmodel does not show how cliques were formed. The ultimate solution to this problem is to consider the weighted adjacency matrix as opposed to the binary adjacency and then construct the distributions of dyads

and higher order interactions.

[WEG2009, p.9]

2.2 RESEARCH ISSUES FOR SOCIAL NETWORKS

Having now laid out the basics of social networks, we would like to raise some research issues. The calculation given above of the two one-mode networks from a two-mode network is dependent on the fact that the networks in question are binary, i.e. the adjacency matrices contain only 0s and 1s. Clearly it would be desirable to allow non

[WEG2009, p.10]

binary entries. Moreover, it seems clear that allowing for the possibility of multi-mode networks is desirable. Another desirable extension is the possibility of allowing the networks to grow or change over time. Two other extensions involve making inferences about missing edges and/or nodes. We address these in turn.

[SHA2008, p.12 cont

....

....

....

....

....

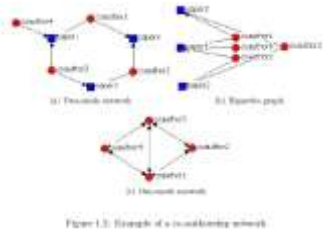
....

....

....

....

....



If we carefully examine the networks in Figures 1.2(a) and 1.3(a) we observe that these

[SHA2008, p.13]

different 2-mode networks in fact have the same 1-mode graphical network representation, ~~see Figures 1.2(c) and 1.3(c).~~ This is due to the fact that when converting to 1-mode some network features are lost;

the same effect when someone projects from 3-D to 2-D. This is an example of how the 1-mode network does not provide sufficient answer of how peer-ties are formed. As a result, the analysis of two-mode networks and one-mode networks should be performed concurrently. …

~~This is important when cliques are present and one needs to determine which members formed which clique. The entrepreneurial and laboratory style of coauthorship networks, which will be discussed in Chapter 5 are different styles, yet the blockmodel of the 1-mode network identifies both as one style.~~

The blockmodel does not show how cliques were formed. The ultimate solution to this problem is to consider the weighted adjacency matrix as opposed to the binary adjacency and then construct the distributions of dyads,

[SHA2008, p.14]

triads, tetrads, pentads, hexads, heptads, and octads.

[WEG2009, p.10 cont]

2.2.1 MULTI-MODE NON-BINARY NETWORKS

The Iraq War gives a perfect example of a situation in which there are obviously more than two classes of actors. At the very least, there are the allied military, the insurgents, and the civilians. Indeed, the allied military could be broken down into U.S. military and other forces. The insurgents could easily be broken into the Shiite militias, the Al-Quaeda insurgents, and the Iranian-sponsored insurgents. The civilians could easily be broken into Shiites, Sunnis, and Christian civilian populations. In the two-mode network case, the two-mode network can be broken into 2! = 2 one-mode networks. This corresponds to the two one-dimensional faces of the two-mode adjacency matrix. Because we have only one unique way of forming the transpose with a two-dimensional matrix, we have only two one-mode networks resulting. Supposing for a moment that we have a three-mode network. Suppose *n, m, k* are respectively the number of actors in each of the three groups. Then these are the dimensions of the sides of the 3D cuboid adjacency matrix. A cuboid matrix is in fact a tensor of rank 3; however, for the purposes of this proposed research discussion, we will use the term cuboid instead.

The 3D cuboid has six faces leading to six different ways to view the block in terms of size, namely, *n* x *m* x *k, n* x *k* x *m, m* x *n* x *k, m* x *k* x *n, k* x*n* x *m* and *k* x *m* x *n*. As a result, the transpose of the 3D cuboid matrix can be done in 3! = 6 different ways. Because the transpose can be done in different ways, the two-mode and one-mode adjacency matrices can be formed in much more complex ways than the simple two-mode social networks. *We propose to explore the three-mode and, more generally, the N-mode social networks through their adjacency tensors of rank 3 and rank N respectively.*

Turning back for a moment from considerations of multimode networks, we focus on non-binary matrices. The reason that the computations for Cmxm = A_{mxn} * A^T_{nxm} and P_{nxn} = A^T_{nxm}* A_{mxn} work in Example 2.1.4.1 is that the matrices are binary. In particular, 1 x 1 = 1 and 0 x 0 = 0 . Of course, if the elements, a_i, are frequency counts or probabilities, this idempotent multiplication will not be the case. It is clear that the multimode case is even more complex. We believe we have an approach based on tensor

[WEG2009, p.11]

decomposition. *We propose to explore methods for addressing conversion of multimode non-binary adjacency tensors and matrices into lower mode networks.*

2.2.2 EVOLVING NETWORKS

Networks evolve over time or at least our knowledge of the actors and their connections evolve over time. Figure 2.3 illustrates the preliminary assessment of connections among the 9/11 hijackers. Figure 2.4 illustrates the assessments of the expanded social network based on additional intelligence information. Of course, it is expected that social networks will evolve as new actors enter the milieu. In addition, of course, while the set of actors in the network itself could be static, the relationships may shift and evolve so that old connections are broken and new one established. The difficulty from a mathematical perspective is that as new actors come into or leave the network, the size of the adjacency matrix or tensor changes.

[SHA2008, p.55]

A cuboid matrix is in fact a tensor of rank 3; however, for the purposes of this research I will use the term cuboid instead.

[SHA2008, p.56]

I would like to discuss how a cuboid is being transposed in 3D. Unlike the 2D rectangular matrix, which only has two faces, the 3D cuboid has six faces leading to six different ways to view the block in terms of size, namely, *n* x *m* x *p, n* x *p* x *m, m* x *n* x *p, m* x *p* x *n, p* x*n* x *m* and *p* x *m* x *n*. As a result, the transpose can be done in six different ways.

[WEG2009, p.11 cont]
Thus the underlying mathematical framework is different. If we view the one-mode adjacency matrix as an operator on a finite dimensional vector space, then as the network evolves, the dimension of the relevant vector space also changes. The implication is that there is no common mathematical framework for the network. The solution would appear to be to assume that the social network has an infinite number of nodes, all but a finite number of the are inactive at any given time. However, they may be activated with null links also being activated. This perspective allows a common infinite-dimensional framework to be in place at all times. Of considerable importance, is the fact that the strength of ties may be time-dependent. This perspective is motivated by our alcohol-modeling in Fairfax County, VA. Not only are new residents coming into the community as well as old ones leaving, but the strength of connections in their multimode network (individuals and alcohol outlets) as measured by conditional probabilities is changing on multiple time scales. The size of this social network is on the scale of 1,000,000 actors. *We propose to develop common mathematical framework for evolving social networks including multimodal networks.*

2.2.3 ESTIMATING MISSING LINKS AND MISSING NODES
Because edges determine connectivity between nodes, they are crucial to the structure of networks and knowing whether or not there is a missing edge in an incompletely observed network is of great importance. In many sampled networks, edges are imperfectly observed because of under-coverage or because actors are intentionally suppressing their roles and linkages to serve different purposes. A clear example of the

[WEG2009, p.12]
latter are networks of terrorists as in Figures 2.3 and 2.4. But criminal networks, networks of spies, even networks of corporations and of countries may want to suppress connections in order to gain strategic advantage. We suggest
a mathematical model to predict unobserved edges and vertices in a network based on covariate information on vertices and edges. The covariates are the exogenous attributes of actors. There are two types of attributes a set of nodes or edges can have, quantitative attributes, which are numerical summaries associated with entities and qualitative attributes, which are categorical summaries associated with entities. Our proposed model consists of two similarity measures calculated simultaneously using both the quantitative and the qualitative attributes derived

exogenously. We note that Marchette and Priebe (2008) develop a method for predicting edges based on a constrained random dot product graph which use endogenous properties of the graph itself.

Our idea is that if two vertices have a high similarity measure, then there is a high probability the vertices have edge connecting them or there is a high potential for forming an edge. Nodes and edges do not necessarily have the same set of attributes. Depending on the network setup and the properties of the entities, different networks may have completely different set of node attributes. Therefore, before applying the proposed method of estimating missing linkages, covariate information needs to be carefully defined.

[SHA2008, p.70]
Estimating Missing Edges And Vertices Using Covariate Information
Because edges determine connectivity between nodes, they are crucial to the structure of networks and knowing whether or not there is a missing edge in an incompletely observed network is of great importance. In many sampled networks, edges are imperfectly observed because of under-coverage or because actors are intentionally suppressing their roles and linkages to serve different purposes.

In this chapter, I present
a mathematical model to predict unobserved edges and vertices in a network based on covariate information on vertices and edges. The covariates are the exogenous attributes of entities. There are two types of attributes a set of nodes or edges can have, quantitative attributes, which are numerical summaries associated with entities and qualitative attributes, which are categorical summaries associated with entities. The model consists of two similarity measures calculated simultaneously using both the quantitative and the qualitative attributes derived

[SHA2008, p.71]
externally as opposed to endogenous approach. In the process of computing the similarity measure between nodes using the quantitative information I use the inner (dot) product technique to obtain an estimate. On the other hand, I use contingency tables and the χ^2 test to obtain another estimate to compute the similarity using qualitative information. The probability of having an edge between two given vertices is then a weighted sum of the two estimates.

If two pairwise vertices wind up having a high similarity measure then there is a high probability the vertices have edge connecting them or there is a high potential for forming an edge. Nodes and edges do not necessarily have the same set of attributes. Depending on the network setup and the properties of the entities, different networks may have completely different set of nodes attributes. Therefore, before applying the method of estimating missing linkages, covariate information need to be carefully defined.

[WEG2009, p.12 cont]
For example, in the author-coauthor social networks, possible attributes on authors and coauthors are *age, education, gender, spoken languages, discipline, number of publications.* However, possible attributes related to papers include *field, topic, keywords, year of publication, publisher, single/multiple author(s)*. In the alcoholconsumer settings, *age, ethnicity, smoker, drug-user, alcoholic, income, job-class* are possible consumers attributes, whereas *zip-code, location, hours-of-day, days-of-week* are some possible attributes associated with alcohol outlets. *We propose to develop methods of inferring the probability of possible missing edges based on inner product and related similarity measures using exogenous attributes.*

Vertices are not less important than edges. In fact, actors are the main element of a network; without actors a network is meaningless. Actors play a significant role in determining the dynamics of a network. We propose
a technique to estimate missing vertices (nodes) in a network. The method is again based on covariate information for vertices (actors) rather than edges, and utilizes the line space of edges which becomes the space of vertices as discussed
above.

In the line space of graphs, vertices become edges and edges become vertices. Consequently, to estimate a missing vertex in the space of

[WEG2009, p.13]
graphs, it suffices to estimate the missing edge corresponding to that vertex in the line space of graphs. In this

regards, we propose to use a mapping to transform from the space of graphs to the line space. Because graphs and matrices are isomorphic (one-to-one and onto), there is a function (transformation) that takes the graph and transforms it from the original space onto the line space and vice versa using matrices. In this sense, the matrix is the operator.

[SHA2008, p.71 cont]
For example, in the author-coauthor social networks, possible attributes on authors and coauthors are *age, education, gender, spoken languages, discipline, number of publications*. However, possible attributes related to papers include *field, topic, keywords, year of publication, publisher, single/multiple author(s)*. In the alcoholconsumer settings, *age, ethnicity, smoker, drug-user, alcoholic, income, job-class* are possible consumers attributes, whereas *zip-code, location, hours-of-day, days-of-week* are some possible attributes associated with ABC stores.

[SHA2008, p.76]
Vertices are not less important than edges. In fact, actors are the main element of a network; without actors a network is meaningless. Actors play a significant role in determining the dynamics of a network. In this section, I will introduce
a technique to estimate missing vertices (nodes) in a network. The method is again based on covariate information for vertices (actors) rather than edges, and utilizes the line space of edges which becomes the space of vertices as discussed
in section (2.9).

~~In optimization theory, maximizing a problem in the dual space is equivalent to; and sometimes tends to be more feasible than, minimizing it in the original space.~~

In the line space of graphs, vertices become edges and edges become vertices. Consequently, to estimate a missing vertex in the space of

graphs, it suffices to estimate the missing edge corresponding to that vertex in the line space of graphs. In this

[SHA2008, p.76]
regards, I use a mapping to transform from the space of graphs to the line space. Because graphs and matrices are isomorphic (one-to-one and onto), there is a function (transformation) that takes the graph and transforms it from the original space onto the line space and vice versa using matrices. In this sense, the matrix is the operator.

[WEG2009, p.14]

Definition 2.10. The *line graph* of $G(V, E)$ also known the *edge graph*, denoted G^l, is a graph satisfying the following criteria:

1. Each vertex of $G^l(V^l, E^l)$ represents an edge of $G(V, E)$.

[WEG2009, p.15]

2. Two vertices v_i^l and v_j^l of $G^l(V^l, E^l)$ are adjacent; i.e. $(v_i^l, v_j^l) \in E^l$ if and only if their corresponding edges are adjacent in $G(V, E)$.

The line graph is intersection graph of the edges of $G(V, E)$, it represents the adjacencies between edges of $G(V, E)$.

van Rooij and Wilf (1965) showed that if G is connected, the sequence $G, G^l, (G^l)^l, \left((G^l)^l\right)^l, \cdots$ of line graphs have four possible behaviors:

1. If G is a cycle graph, C_n, then $G^l$and each subsequent line graph is isomorphic to G itself. Cyclic graphs are the only connected graphs for which $G^l = G$.
2. If G is a claw $K_{1,3}$, then G^l and all subsequent line graphs are C_3.
3. If G is a path graph P_n, then each subsequent line graph is a shorter path P_{n-1} until eventually P_0 terminates with an empty graph.
4. In all remaining cases, the sizes of the line graphs increase without bound.

The idea then is to convert G(V, E) into $G^l(V^l, E^l)$, and then apply the method for inferring edges to $G^l(V^l, E^l)$. *We propose to investigate the feasibility of inferring missing nodes by using the duality principle of line graphs and the procedures for inferring missing edges described above.*

[SHA2008, p.67]

2.9 Line Graphs

Definition 2.9.1. The line graph of $G(V, E)$ also known the edge graph, denoted G^l, is a graph satisfying the following criteria:

1. Each vertex of $G^l\left(V^l, E^l\right)$ represents an edge of $G(V, E)$.
2. Two vertices v_i^l and v_j^l of $G^l\left(V^l, E^l\right)$ are adjacent; i.e. $\left(v_i^l, v_j^l\right) \in E^l$ if and only if their corresponding edges are adjacent in $G(V, E)$.

The line graph is intersection graph of the edges of $G(V, E)$, it represents the adjacencies between edges of $G(V, E)$.

van Rooij and Wilf (1965) showed that if G is connected the sequence $G, G^l, \left(G^l\right)^l, \left(\left(G^l\right)^l\right)^l$ of line graphs have four possible behaviors:

1. If G is a cycle graph C_n then G^l and each subsequent line graph is isomorphic to G itself. Cyclic graphs are the only connected graphs for which $G^l \cong G$.
2. If G is a claw $K_{1,3}$, then G^l and all subsequent line graphs are C_3.
3. If G is a path graph P_n, then each subsequent line graph is a shorter path P_{n-1} until eventually P_0 terminates with an empty graph.
4. In all remaining cases, the sizes of the line graphs increase without bound.

[WEG2009, p.15 cont]

3 COMPUTER NETWORKS

The concepts that are applied to social networks can also be applied to computer networks. If the computers (or routers) are nodes (actors), then their edges correspond to direct connections. Thus the node degree of a computer (or router) is just the number of other computers directly connected to it. Computer networks are generally designed (or perhaps not designed) on an evolving basis, with new connections being made on a greedy basis. Like human social networks computer networks tend to be scale free

A scale-free network is a network whose degree distribution follows a power law. The fraction *P(k)* of nodes in the network having *k* connections to other nodes is, for large values of *k*, P(k) ~$k^{-\gamma}$ where γ is a constant whose value is typically in the range $2 < \gamma < 3$This means a few computers (nodes) will have a high node degree, most likely servers, while many computers will have a low node degree, most likely clients. It is our premise that neither situation is particularly good from a network security perspective. A very high degree node, a node having high *centrality* and especially

[WEG2009, p.16]

betweenness centrality, is vulnerable because it is likely a target of a *denial-of-service attack*. On the other hand, a node with low degree is unimportant in the network could easily be the target of a *trojan* and could be made to function as a *zombie*. A low degree node is unlikely to have much attention from system administrators, and consequently could be spewing out spam email or worse yet private and classified information. Finally an unimportant node could be used by a threatening insider.

3.2 NETWORK ASSESSMENT METRICS

We describe and use the following metrics in assessing and improving network robustness and efficiency:

3.2.1 Average Shortest Path Length (ASPL) - ASPL can be used as an indicator of the efficiency of the network and is the average of all of the shortest paths from every node to the other connected nodes. This means that the smaller the ASPL, the more efficient the network may be.

3.2.2 Network Diameter (ND) - Network diameter can be used as an indicator of the efficiency factors of a network and is the longest of all of its shortest paths. This could be defined as the "maximum of all of the minimum paths" (MaxMin).

Generally, the size of a network's diameter is indicative of how spread out the network may be, and the more spread out the network is, potentially, the less efficient it may be.

3.2.3 Average Node Degree (AND) and Average Node Degree Squared (ANDS) - ANDS represents the average of the squared values of the degrees of all of the nodes on the network. Typically, one would use the average of the node degrees as an indicator. This value, however, will never vary if the total number of links on the network (Network Link Budget) remains constant. As this will be the case in some of our proposed research and evaluations, we opt to use the squared values of the node degrees as opposed to the actual node degrees.

[REZ2009, p46]

A scale-free network is a network whose degree distribution follows a power law. The fraction *P(k)* of nodes in the network having *k* connections to other nodes is, for large values of *k*, P(k) ~$k^{-\gamma}$ where γ is a constant whose value is typically in the range $2 < \gamma < 3$.

[REZ2009, p35]

3.4. Network Assessment Metrics

Based on these definitions, I describe and use the following metrics in assessing and improving network robustness and efficiency:

• Average Shortest Path Length (ASPL) – ASPL can be used as an indicator of the efficiency of the network and is the average of all of the shortest paths from every node to the other connected nodes. This means that the smaller the ASPL, the more efficient the network may be.

…(text and equation)…

• Network Diameter (ND) – Network diameter can be used as an indicator of the efficiency factors of a network and is the longest of all of its shortest paths. This can be defined as the "maximum of all of the minimum paths" (MaxMin):

ND = Max [Min (all SPLs)]

Generally, the size of a network's diameter is indicative of how spread out the network may be, and the more spread out the network is, potentially, the less efficient it may be.

• ~~Degree Standard Deviation (DSD~~) *(text+equation) this just was reordered elsewhere*

• Average Node Degree (AND) and Average Node Degree Squared (ANDS) – ANDS represents the average of the squared values of the degrees of all of the nodes in the network. Typically, one would use the Average Node Degree (AND) as ~~a measurement or~~ an indicator. This value, however, will never vary if the total number of links and nodes on the network (Network Link Budget) remains constant. As this is the case in most of the research in this dissertation, the AND will not be a useful indicator. As a result, I opted to use the squared values of the node degrees as opposed to the actual node degrees in order to be able to capture differences in the node degree centralizations between a given network and a revised instance of such a network: ANDS = SSND / *N* where SSND is the Sum of the Squares of the Node Degrees and *N* is the total number of nodes on the network

[WEG2009, p.16 cont]
Large values of ANDS may be indicative that the network is vulnerable, since the network may take a significant hit if the nodes with high connectivity were to be removed or Lost. Large values of ANDS could also be an indication that the network is "over-connected" and there are excess links, and therefore it is not a very efficient network. As a result, ANDS may be used as a criterion for robustness, as well as one for efficiency of networks.

[WEG2009, p.17]
3.2.4 Degree Standard Deviation (DSD) - DSD represents the standard deviation of all the degrees of all the nodes in the network and is an indicative of variability of the degree densities among the nodes.

3.2.5 Average Node Betweenness (ANB) - ANB represents the average (mean) of the betweenness values for all of the nodes on the network. Node Betweenness represents the number of times a node appears in the set of all the shortest paths that connect all of the nodes of the network.

3.2.6 Highest Link Betweenness (HLB) - HLB represents the highest value of the betweenness from among the betweenness values for all of the links on the network.

3.2.7 Network Degree Centralization (NDC) - Degree centrality measures how 'concentrated' the degree centralities of the actors are in the network.

3.2.8 Network Closeness Centrality Mean (NCCM) - NCCM represents the mean of the closeness values for all of the nodes on the network.

3.2.9 Average Node-Based Disconnection Ratio (ANBDR) -Node-Based Disconnection Ratio measures the ratio of the network disconnection caused by the removal of a node on the network. ANBDR is the average of all of the disconnection ratios caused by removing a node on the network
This is a measure of robustness of the network.

3.2.10 Average Link-Based Disconnection Ratio (ALBDR) - Link-Based Disconnection Ratio measures the ratio of the network disconnection caused by the removal of a link on the network. ALBDR is the average of all of the disconnection ratios caused by removing a link on the network.
Similarly, this is also a measure of robustness.

3.3 NETWORK OPTIMIZATION
As indicated above, our premise is that nodes that are too central are vulnerable and thus make the network vulnerable. Likewise, nodes that are of low centrality are also

[WEG2009, p.18, 19, 20, most of 21 omitted: no antecedents found, via quick searches in Epstein&Axtell(1996) and North&Macal(2007). The writing is choppy, but Wegman should have known this material, and the proposal was written fairly quickly.]

[REZ2009, p37]
Large values of ANDS may be indicative that the network is vulnerable, since the network may take a significant hit if the nodes with high ~~volume of~~ connectivity were to be removed or lost. Large values of ANDS could also be an indication that the network is "over-connected" and there are excess links, and therefore it is not a very efficient network. As a result, ANDS may be used as a criterion for robustness, as well as one for efficiency of networks.

[REZ2009, p35] *(this was just re-ordered)*
• Degree Standard Deviation (DSD) – DSD represents the standard deviation of all the degrees of all the nodes in the network and is an indicative of variability of the degree densities among the nodes: *(equation)*
[REZ2009, p37]
• Average Node Betweenness (ANB) – ANB represents the average of the betweenness values for all of the nodes in the network

• Highest Link Betweenness (HLB) – HLB is the largest of the betweenness values for all of the links in the network.

• Network Degree Centralization (NDC) – Degree centrality measures how 'concentrated the degree centralities of the actors are in the network.

• Network Closeness Centrality Mean (NCCM) – NCCM represents the mean of the closeness values for all of the nodes in the network
[REZ2009, p38]
• Average Node-Based Disconnection Ratio (ANBDR) – Node-Based Disconnection Ratio measures the ratio of the network disconnection caused by the removal of a node on the network. ANBDR is the average of all of the disconnection ratios caused by removing a node on the network.

• Average Link-Based Disconnection Ratio (ALBDR) –Link-Based Disconnection Ratio measures the ratio of the network disconnection caused by the removal of a link on the network. ALBDR is the average of all of the disconnection ratios caused by removing a link on the network.

*The use of "**actors**" is odd here, although Wegman and students have elsewhere sometimes applied SNA terms to computer networks in place of standard network terms, such as nodes.*

[WEG2009, p.21]

An agent-based model (ABM) is a computational model for simulating the actions and interactions of autonomous individuals in a network, with a view to assessing their
[WEG2009, p.22]
effects on the system as a whole. It combines elements of game theory, complex systems, computational sociology, multi-agent systems, and evolutionary programming. Monte Carlo Methods are used to introduce randomness.
Two informative books in the area are North and Macal (2007) and Epstein and Axtell (1996). The latter book was among the very first large-scale attempt to create agent-based models.

The next section likely comes originally from Brush(1967) or Binder(2001), or somewhere else, but I did find a copy that precedes [WEG2009], so it is unlikely that Wegman wrote this. The inclusion of unquoted text is certainly consistent. Axtell was on Sharabati's committee, so that may be a connection.

[WEG2009, p.22 (most), 23, 24 (most) skipped, no antecedents seen]

[WEG2009, p.24] ***Very likely original(s) were Brush or Binder.***
5.4 SUMMARY
Agent-based modeling does not have a strong mathematical and statistical foundation. One interesting connection to agent-based modeling is the Ising model, which has a fairly well-developed mathematical/statistical underpinning.
The Ising model, named after the physicist Ernst Ising, is a mathematical model in statistical mechanics. See Brush (1967) and Binder (2001).
It has since been used to model diverse phenomena in which bits of information, interacting in pairs, produce collective effects.

The Ising model is defined
[WEG2009, p.25]
on a discrete collection of elements (agents) with variables called spins, which can take on the value 1 or -1. The spins, s_i, interact in pairs, with energy that has one value when the two spins are the same, and a second value when the two spins are different.

While the Ising model is a simplified microscopic description of ferromagnetism, it is still extremely important because of the universality of the continuum limit. Universality means that the fluctuations near the phase transition are described by a continuum field with a free energy or Lagrangian which is a function of the field values. Just as there are many ways to discretize a differential equation, all of which give the same answer when the lattice spacing is small, there are many different discrete models that have the exact same critical behavior, because they have the same continuum limit.

Wikipedia, Agent-based model, 10/31/07 (first appearance of text)
en.wikipedia.org/w/index.php?title=Agent-based_model&oldid=168343713
26 September 2008
en.wikipedia.org/w/index.php?title=Agent-based_model&oldid=241095322 09/26/08
An agent-based model (ABM) is a computational model for simulating the actions and interactions of autonomous individuals in a network, with a view to assessing their

effects on the system as a whole. It combines elements of game theory, complex systems, emergence, computational sociology, multi agent systems, and evolutionary programming. Monte Carlo Methods are used to introduce randomness. …
Joshua M. Epstein and Robert Axtell developed the first large-scale ABM, the Sugarscape, to simulate and explore the role of social phenomenon such as seasonal migrations, pollution, sexual reproduction, combat, and transmission of disease and even culture.

It is possible that the beginning words were picked up from something Wegman and/or Said wrote, but more likely, they were derived from someone well-published in the ABM field. By late 2008, Said and Wegman had written some ABM papers, so history of antecedents in unclear. This is minor, but odd.

English text at Chinese website, 12/30/08
zh-yue.wikipedia.org/w/index.php?title=User:Hillgentleman/i_sing&oldid=234927

The Ising model, named after the physicist Ernst Ising, is a mathematical model in statistical mechanics.
It has since been used to model diverse phenomena in which bits of information, interacting in pairs, produce collective effects.

en.wikipedia.org/w/index.php?title=Ernst_Ising&oldid=273634420 02/27/09[74]
The Ising model is defined

on a discrete collection of elements (agents) with variables called spins, which can take on the value 1 or -1. The spins, s_i, interact in pairs, with energy that has one value when the two spins are the same, and a second value when the two spins are different.
English text at Chinese website, 12/30/08
While the Ising model is a simplified microscopic description of ferromagnetism, it is still extremely important because of the *universality* of the continuum limit. Universality means that the fluctuations near the phase transition are described by a continuum field with a free energy or Lagrangian which is a function of the field values. Just as there are many ways to discretize a differential equation, all of which give the same answer when the lattice spacing is small, there are many different discrete models that have the exact same critical behavior, because they have the same continuum limit.

[74] Of course, this followed [WEG2009], but *likely* was copied from elsewhere.

[WEG2009, p.25 cont]

The Ising model undergoes a phase transition between an ordered and a disordered phase in 2 dimensions or more. In 2 dimensions, the Ising model has a strong/weak duality (between high temperatures and low ones) called the Kramers-Wannier duality. The fixed point of this duality is at the second-order phase transition temperature.

The experimentally observed critical fluctuations of ferromagnets near the Curie point and of fluids at the vapor/liquid critical point are described exactly by the critical fluctuations of the Ising model. The same is true for the simplest statistical models in three dimensions whose fluctuations can be described by a single scalar field, the local magnetization in a near-critical magnet or the local density in a near-critical fluid. All these systems have fluctuating clusters whose fractal scaling laws and long distance correlation functions are quantitatively predicted by the model. Apart from the continuum limit, many discrete systems can be mapped exactly or approximately to the Ising system. The grand canonical ensemble formulation of the lattice gas model, for example, can be mapped exactly to the canonical ensemble formulation of the Ising model. The mapping allows one to exploit simulation and analytical results of the Ising model to answer questions about the related models.
The Ising model can be thought of as a Markov random field on a square grid, where the maximal graph cliques are edges (i.e. pairs of neighboring vertices).

While a very simplified version of an ABM, the Ising model clearly has elements that have local interactions based on rules for the spins and which have emergent behavior on a macroscale. The Ising model is generally constructed on a lattice. So it is an extremely simplified version of what might be thought of as an ABM. Nonetheless, it is suggestive
[WEG2009, p.26]
that a more sophisticated mathematical/statistical framework could be developed for agent-based models.
We propose to develop the mathematical/statistical framework of agent-based models operating in a social network environment. Agent-based models do not have a strong mathematical foundation. However, in analogy to the Ising model, we propose to investigate the limiting behavior of agent-based models on a lattice as the lattice size shrinks.

[WEG2009, p.26 (rest), p.27 skipped –summary of proposed tasks]

The entire section on Ising models seems very strange, out of place with the rest of the proposal, which was related to work that had been done by Wegman's students. This text seems very unlikely to have been written originally by Wegman or his students.

English text at Chinese website, 12/30/08

The Ising model undergoes a phase transition between an ordered and a disordered phase in 2 dimensions or more. In 2 dimensions, the Ising model has a strong/weak duality (between high temperatures and low ones) called the Kramers-Wannier duality. The fixed point of this duality is at the second-order phase transition temperature.

The experimentally observed critical fluctuations of ferromagnets near the Curie point and of fluids at the vapor/liquid critical point are described exactly by the critical fluctuations of the Ising model. The same is true for the simplest statistical models in three dimensions whose fluctuations can be described by a single scalar field, the local magnetization in a near-critical magnet or the local density in a near-critical fluid. All these systems have fluctuating clusters whose fractal scaling laws and long distance correlation functions are quantitatively predicted by the model. Apart from the continuum limit, many discrete systems can be mapped exactly or approximately to the Ising system. The grand canonical ensemble formulation of the lattice gas model, for example, can be mapped exactly to the canonical ensemble formulation of the Ising model. The mapping allows one to exploit simulation and analytical results of the Ising model to answer questions about the related models.
The Ising model can be thought of as a Markov random field on a square grid, where the maximal graph cliques are edges (*i.e.* pairs of neighboring vertices).

The End