SCM System

Joel H. Levine*, Kathleen M. Carley

June 3, 2016
CMU-ISR-16-108

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Center for the Computational Analysis of Social and Organizational Systems
CASOS technical report.

"Measure what is measurable, and make measurable what is not so." - Galileo Galilei

Quoted in I Gordonand and S Sorkin, The Armchair Science Reader *(New York 1959).* ***Quotations by Galileo Galilei***
http://www-history.mcs.st-and.ac.uk/Quotations/Galileo.html

This work was supported in part by the Office of Naval Research (ONR) N000141512797 Minerva award for Dynamic Statistical Network Informatics, and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research or the U.S. government.

*Joel H. Levine is a Professor at Dartmouth College, Hanover, NH

Carnegie Mellon

ABSTRACT

Socio-cultural cognitive maps (SCMs) are the best-fit network model to the set of underlying node to variable data. SCM's permit objective visualization of the network, inference about the impact of changes in the underlying conditions influencing the nodes, and comparison of disparate data. This report details the process of creating and assessing these SCMs. First a general introduction is provided and then a step by step guide based on a cognitive walkthrough is presented.

INTRODUCTION

How do we make sense of communities? How do we understand and predict changes in these communities? From a socio-cultural perspective addressing these questions means attaining a structural understanding of actors, issues, and the relations connecting them. Or in other words, it means answering these questions: 1) Who are the critical actors, particularly the political, tribal, religious, economic, educational and religious elite and associated groups. 2) On what specific micro-issues are the interests of these elites and their groups aligned, and on what issues do they compete? 3) What is the basis for those relations, alliances and conflicts e.g., are they based in economics, status, education, religion, or location. Further, it is important to not only understand the lay of the land, but to use that information to assess the community of actors of interest vis-à-vis some issue, e.g. resilience, cyber-attacks, or deterrence given those relations and the basis for them. And 4) How will the community change its position on a broad area of concern, an issue, given changes in the alliances and competitions on the more micro issues, perhaps due to the underlying basis for a relation being altered, or an actor being removed? That is, how stable is the community, how resilient is the community, given change at the actor level or basis for alliance/competition level?

We suggest that these questions can be addressed through the development and assessment of socio-cognitive cultural maps (SCMs). We further suggest that it is critical to develop, visualize and assess these SCMs quickly, and in a fashion that supports 'what-if' reasoning. The SCM is the best-fit model of these underlying relations among actors in the region of interest based on a socio-cognitive understanding of the social and cultural similarities and differences among a *community* of *actors* given a set of *topics* relevant to an *issue*.

In general, in an SCM the actors might be individuals or collectives and the set of actors in the SCM are the "*community*". This *community* may be comprised of individual *actors* that are public persona (e.g., political elite such as heads of country), groups (e.g., ethno-religious, socio-economic, covert and political groups), nation states or governance bodies (e.g., the UN), or key stakeholder groups (e.g., the executive or military branch of a country). An SCM is typically developed around an *issue*. This *issue* is the thing about which the analyst wants to know the community's current position, and how that position will change if the set of actors, the relations among them, or the basis of those relations changes. Illustrative *issues* are the resilience of the community of groups within a country to changes in socio-economic conditions such as changes in wealth and education; or the danger of a nuclear event (deterrence) in a region of interest given changes in the force posture of key stakeholders; or the resiliency to cyber-attacks of third-world countries given the global cyber-threat mapping. For each *actor* in the *community*, pursuant to these *issues*, there are a number of *topics* of relevance on which the individual *actors* have "scores." These topics include issues, beliefs, or norms where the actor has a position (or score) such as the belief that an ally will support them in the event of a nuclear incident or the level of concern with climate change or a general socio-demographic attribute such as level of education or wealth, or an infrastructure attribute such as internet penetration. These *topics* are the dimensions along which actors can be similar or different. The SCM

can be represented as a network where the nodes are *actors* and the links express the connectivity among the *actors* taking into account either similarity and dissimilarity of the two actors given *issue* relevant *topics* and/or *issue* relevant networks (such as trade volume networks, hostility networks, and alliance agreements).

Mathematically, the SCM is a reduction of the more complex detail available in the hyper-cube where the dimensions are actors by actors by topics by topics; the actor-topic links are the strength of connectivity; the topic-topic-links are the co-presence or covariance of the topics; and the actor-actor links are inferred from the other dimensions given the degree to which two actors share the same topics and the extent of the connectivity between those topics. Finding the SCM and assessing it, however, is a time intensive process that requires the analyst to make a large number of choices regarding the underlying data. The goal then is to develop an SCM technology that supports the a) rapid development of SCMs, b) is sensitive to cultural differences, c) results in an interpretable model, and d) is usable for assessing possible interventions. The algorithms needed for generating, assessing, and visualizing SCMs need to be robust, scalable, reusable, and reproducible. However, there is no such technology. In contrast, in this document we lay out a possible technology and walk the reader through the underlying steps in the construction, visualization and use of SCMs.

This document is organized as follows. We begin by describing the vision of SCMs and the role of linear methods in that process. Then we describe the cognitive walkthrough process used to identify the workflow and technologies needed to create, use, visualize and assess SCMs.

LINEAR METHODS FOR NOT-YET-MEASURED VARIABLES

The linear model is the best tool we have for describing the relation between two variables. It says that variable *Y* is a linear function of variable *X*. It is simple and powerful, yet limited. It does not apply to categorical variables nor to networks that have structures but no variables. It is limited to numerical variables.

We can and have developed other tools for correlations among non-numerical objects, but there is an alternative — which is to extend number and the linear model to these objects. We do that by assuming that numbers and linear relations exist for these variables but have not yet been discovered. Then we attempt to reverse engineer these not-yet-measured variables and validate the assumptions under which they have been discovered.

There is ample reason to suspect that these numbers and variables exist. For example, Figure 1 shows data from the *Washington Post*, describing prior activities of terrorists involved in the 911 attack on the United States. The data tell us that Nawaf al-Hazmi (column 3) and Khalid al-Mindhar (column 4) appeared together in a video (row 1), that Mohammed Atta and Marwan al-Shehhi took flying lessons in Venice Florida (row 3), and so on for 31 activities.

Joint Events in Order by Date		Original (Reformatted, in Excel, from The Washington Post, Sunday, October 21, 2001, Sources: Staff research and news Accounts, Chart by Bill Webster, Washington Post)	Pentagon					World Trade Center 1						World Trade Center 2						Pennsylvania			
			Hanjour (Pent)	Moqed (Pent)	Nawaf Alhazmi (Pent)	Salem Alhazmi (Pent)	Almihdhar (Pent)	Atta (WTC1)	Waleed Alshehri (WTC1)	Wail Alsheri (WTC1)	Suqami (WTC1)	Alomari (WTC1)	Al-Shehhi (WTC2)	Banihammad (WTC2)	Ahmed Alghamdi (WTC2)	Hamza Alghamdi (WTC2)	Mohand Alsheri (WTC2)	Jarrah (PA)	Alnami (PA)	Saeed Alghamdi (PA)			
2000	1	Video w/ bin Laden Operatives	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0			
	2	Flying lessons San Diego/ Move in w/ Shalk	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0			
	3	OK flight school, no enroll / pilot training Venice F	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0			
	4	jet simulator lessons Miami Fla	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0			
2001	5	rent mailbox in Del Ray Beach Fla	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0			
Jan-June	6	visit Atlanta rent Piper Cherokee plane	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0			
	7	leave Hamburg apartment	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0			
	8	get driver's licences Coral Springs, Fla	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0			
	9	together Hamlet Country Club Del Ray Beach Fla	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0			
	10	rent Delray Racquet Club	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1			
2001	11	World Gym Del Ray Beach and Boynton Beach	0	0	0	0	0	1	1	1	1	0	1	1	0	0	0	0	0	0			
July-Septembe	12	arrive JFK from Saudi Arabia	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0			
	13	Car rental Wayne NJ (rental 2)	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
	14	VA drivers licenses and ID cards	1	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0			
	15	Car rental Wayne NJ	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
	16	One-way tickets via net UA 175	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0			
	17	Check-in Deerfield Beach, buy tix via net	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0			
	18	One-way tickets via net	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0			
	19	Internet Ticket Activity	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0			
	20	Buy Tix using same address and Visa	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
	21	move out of NJ appt.	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
	22	Internet Ticket Activity and Hotel check-out	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	1	0	0			
	23	Together in Valencia Motel	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0			
	24	gym passes purchased	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0			
	25	one way tix bought in Lauderdale	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1			
	26	Ft. Lauderdale to Newark	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1			
	27	raw bar in Hollywood Fla/check out DFld. Beach	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0			
	28	buy tickets at BWI	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0			
	29	hotel in downtown Boston	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	0	0			
	30	Days Inn outside Boston	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0			
	31	Comfort Inn S.Portland Maine	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0			

Figure 1 Activities of the 911 Terrorists - Washington Post Data Ordered by Date and Target

Is there order in these data? Is there a dimension? The *Post* presented the data by date and by airplane, but re-organizing the data as in Figure 2 strongly suggests there is a linear order. Visual patterns prove nothing but they can suggest a great deal, suggesting an order and intervals that exist but have not been measured.

Joint Events index ed by Date	Levine P/A Pattern Ordering. (described in "Joint-Space Analysis of 'Pick/Any' Data: Analysis of Choices from an Unconstrained Set of Alternatives," Psychometrika, March, 1979.) © 11/5/01 joel.levine@Dartmouth.edu	Nawaf Alhazmi (Pent)	Almihdhar (Pent)	Moqed (Pent)	Salem Alhazmi (Pent)	Hanjour (Pent)	Atta (WTC1)	Al-Shehhi (WTC2)	Waleed Alshehri (WTC1)	Wail Alsheri (WTC1)	Suqami (WTC1)	Alomari (WTC1)	Jarrah (PA)	Banihammad (WTC2)	Mohand Alsheri (WTC2)	Ahmed Alghamdi (WTC2)	Hamza Alghamdi (WTC2)	Alnami (PA)	Saeed Alghamdi (PA)	
1	Video w/ bin Laden Operatives	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
2	Flying lessons San Diego/ Move in w/ Shalk	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
12	arrive JFK from Saudi Arabia	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
28	buy tickets at BWI	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
20	Buy Tix using same address and Visa	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
24	gym passes purchased	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0		
23	Together in Valencia Motel	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0		
19	Internet Ticket Activity	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0		
21	move out of NJ appt.	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0		
13	Car rental Wayne NJ (rental 2)	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0		
15	Car rental Wayne NJ	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0		
3	OK flight school, no enroll / pilot training Venice Fla	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0		
4	jet simulator lessons Miami Fla	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0		
6	visit Atlanta rent Piper Cherokee plane	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0		
7	leave Hamburg apartment	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0		
9	together Hamlet Country Club Del Ray Beach Fla	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0		
27	raw bar in Hollywood Fla/check out DFld. Beach	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0		
17	Check-in Deerfield Beach, buy tix via net	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0		
31	Comfort Inn S.Portland Maine	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0		
22	Internet Ticket Activity and Hotel check-out	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0		
11	World Gym Del Ray Beach and Boynton Beach	0	0	0	0	0	1	1	1	1	1	0	0	1	0	0	0	0	0		
8	get driver's licences Coral Springs, Fla	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0		
29	hotel in downtown Boston	0	0	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0	0		
14	VA drivers licenses and ID cards	0	0	1	1	1	0	0	0	0	0	1	0	0	0	1	0	0	0		
16	One-way tickets via net UA 175	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0		
5	rent mailbox in Del Ray Beach Fla	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0		
18	One-way tickets via net	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0		
30	Days Inn outside Boston	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0		
10	rent Delray Racquet Club	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1		
25	one way tix bought in Lauderdale	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1		
26	Ft. Lauderdale to Newark	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1		

Figure 2 : Activities of the 911 Terrorists -Washington Post Data — Reorganized

Subject to test, the discovery of the missing numbers begins by practicing on data for which the numbers are known and by learning the rules by which the known x's and y's are linked to the fine detail of the data. Then, having learned the rules, we switch to more challenging data in which the x's and y's are unknown and work backward from the fine detail of the data to estimates of the not-previously-measured x's and y's — assuming and testing the assumption that the rules continue to apply. The SCM process is a human-in-the-loop semi-automated approach to doing this test and discovery, to finding the patterns hidden in, but not previously measured from, the raw data.

Now let's move to another set of data referring to the height and weight of individuals. In Figure 3, the height-weight data demonstrate the empirical link between the numbers for height and weight (shown at the left and at the bottom) and the data for joint frequencies of heights and weights (shown in the cells). We can ask of these data: What are the rules that govern this empirical link?

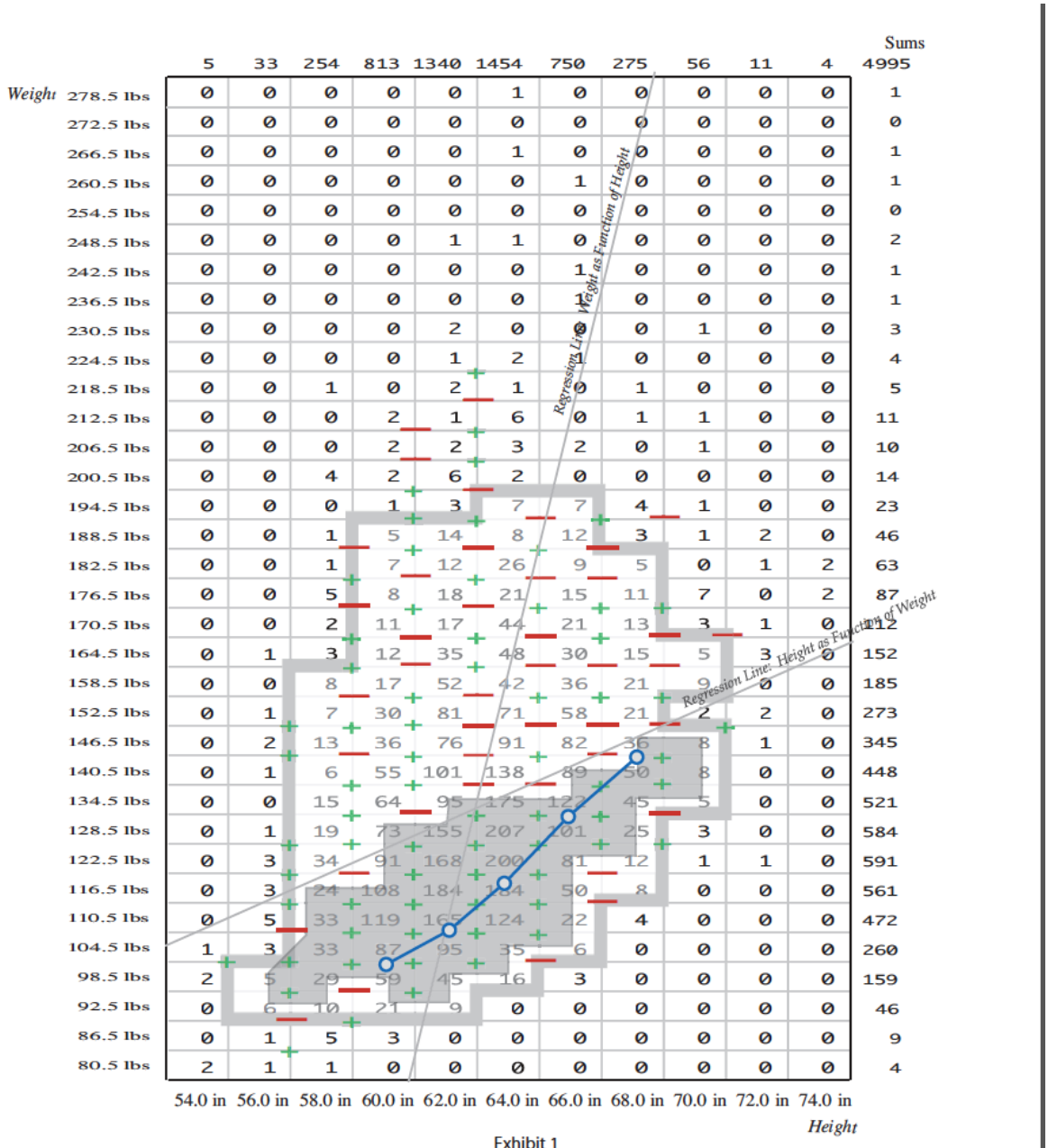

Exhibit 1
Height by Weight

Data reporting height and weight for 4,9995 women, Great Britain, 1951. From Kendall and Stuart's The Advanced Theory of Statistics, *Volume 2, 4th Edition, 1979, p. 300. Original source, "Women's Measurements and Sizes," H.M.S.O., 1957. Cell counts are the numbers of women reporting that combination of column-height and row weight. Pluses and minuses are located at the centers of four cell subsets They report the sign of the log odds ratio within the set. Minuses mark combinations in which odds indicate lighter weight for taller women. The grey outline sets off the region of log odds ratios based on relatively large counted values, greater than or equal to five. The objective line is indicated by the circles.*

Figure 3. Image of relation between Height and Weight

The do's and the don'ts

The dont's

In principle, this might be simple. We might assume that the data have a two-variable "normal" (Gaussian) distribution, and use the Gaussian assumption as the rule, acknowledging that the Gaussian assumption might be only an approximation.

The problem is that, in so doing, we assume-away what may be (and is) a real structure in the data, a structure that is both not Gaussian and rich with information: We are not free to assume rules according to convenience or convention: The "Rules" linking numbers to data are theories and need to be respected as such.

How do we know it is wrong? The goodness (or badness)of fit of the Gaussian assumption to these data can be observed by writing and testing 343 simultaneous equations that link the known heights and weights to the 343 known frequencies in the data — where the ordinates of the Gaussian should be approximately proportional to the observed frequencies. By direct computation, we know the 5 parameters of the Gaussian that are required by the equations (2 means, 2 standard deviations and one correlation, r). And we can the constant of proportionality, a sixth parameter, by chi-square best fit between the equations and the data. The result is a chi-square error of 879,401 with 368 degrees of freedom:

Equation 1

$$\begin{cases} F(i,j) \propto \dfrac{1}{2\pi s_x s_{y\sqrt{1-r^2}}} exp\left(-\dfrac{1}{2(1-r^2)}\left[{X_i}^2 - 2rX_iY_j + {Y_j}^2\right]\right), for\ all\ rows\ i\ and\ columns\ j \end{cases}$$

where $X_i = \frac{x_i - \bar{x}}{s_x}$ and $Y_j = \frac{y_j - \bar{y}}{s_y}$

where $\bar{x}$, $\bar{y}$, s_x, s_y, and r are the appropriate means, standard deviations, and correlation,

and where the equation is approximate because the bivariate normal is continuous while the frequencies are constructed by grouping the data into intervals,

That is a bad fit: Using the theoretical properties of chi-square as a convention, the chi-square error of a good fit should be in the neighborhood of 368 (the number of degrees of freedom). By contrast, the chi-square value of 879,401 (associated with the normal) is about 2,000 greater (worse than) this target.

Assessing the goodness of fit by a different criterion, the error (associated with the Gaussian) can be compared to the error from a model we know to be false: comparing the error of the normal to the error associated with the obviously false assumption that there is no correlation. The best-fit no-correlation model produces a chi-square error of 1,127, with 330 degrees of freedom: This means that the assumptions embedded in the Gaussian are not only inappropriate but worse than the assumption that there is no-correlation at all – where the null reduces error to about one-tenth of one percent of the error associated with the Gaussian.

More important for a scientist, while the data are not Gaussian but they do show a pattern. For example, consider the isolated data for women at 62 and 64 inches with weights 146.5 to 152.5 pounds. In this subset of the data the *taller* women are lighter: The odds that a 62 inch women will weigh 152.5 lbs, as compared to 146.5, are 81 to 76, approximately 1.06 to 1. By contrast, at 71 to 91, the odds that the taller women will have the heavier weight are smaller .

152.5 lbs	81	71
146.5 lbs	76	91
	62.0 in	***64.0 in***

Figure 4. Taller women are lighter – data subset.

This reverse correlation is both counter-intuitive and impossible — if the data were Gaussian. Yet it occurs in roughly one third of the subsets that can be isolated (even excluding low-frequency data by using only those examples that show a minimum of 5 people per cell.) [See Levine 1995?]

The message, is that the reverse local correlations are an unexpected but real and orderly property of the data.

And the do's

Reducing the algebra and the specificity) of the Gaussian, the better model replaces the key factor (the correlation factor) with an expression $|X-Y|^a$ where the Gaussian uses the expression $|X-Y|^2$. And it drops assumptions about the one-variable terms of the Gaussian, other than to assume that they are multiplicative and can be replaced with multiplicative parameters to be estimated from the data. See Figure 5 for an illustration.

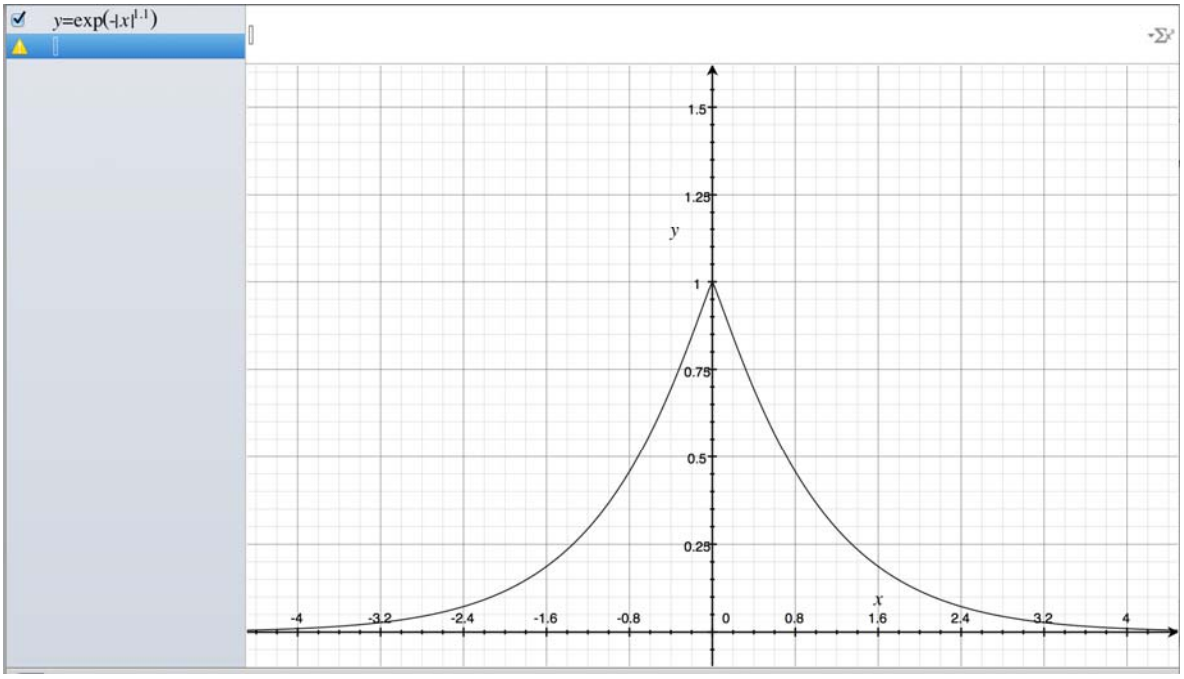

Figure 5: Illustrative Non-Gaussian, $a \approx 1.1$

It tells us that there exist strongly preferred combinations of height and weight, more narrowly limited by the data than the averages or a bell-shaped (normal) distribution.

It sends research down a different path: Where Gaussians variables are thought to be generated by aggregations of many uncorrelated causal variables, the "spike" is not Gaussian. Fitting these "spiked" distributions to the data, fits the data (chi-square ≈ 248) but estimates a substantially different non-Gaussian line, estimating a linear slope of 4.3 lbs per inch (as compared to the to 2.6 lbs per inch by standard regression or the 8.4 pounds per inch obtained by attempting to fit the full two-variable Gaussian). And unlike the conventional estimates of the number of pounds per inch, the non-Gaussian (non-least squares) hypotheses is backed up by a tight fit to the data.

Making no *a priori* decision about *a*, the improved model replaces *2* with a value to be estimated from the data.

$$F(i,j) = R_i \times C_j \times 2^{-d_{ij}{}^a} \qquad (4)$$

$$= \boldsymbol{X\text{–}factor} \times \boldsymbol{Y\text{–}factor} \times \boldsymbol{X\&Y\ linear\ correlation\ factor}$$

where the "All factor" (of Equation 2) has been absorbed into the X-Factor and Y-factor,

where $d_{ij} = |X_i - Y_j|$,

$X_i = (x_i - M_x)/W_x$ and $Y = (y - M_y)/W_y$,

where base *"2"* is used in preference to "*e*" for convenience with applications (without affecting the goodness of fit),

where the *M* and *W* symbols indicate that these additive and multiplicative parameters of the linear relations need not be means and standard deviations,

and where *a* (estimated at 1.1 for the height-weight data) is an attenuation constant that governs the rate of descent (of the correlation factor) with respect the to distance of each combination of X and Y from the line X = Y.

With this not-necessarily quadratic supplement to the kinds of correlation that will be recognized, the χ^2 best fit of *a* is 1.1 which reduces the error another 63% (another 2.7-fold) to 248 (with 325 degrees of freedom). With this rule/theory the magnitude of the chi-square is now less than the magnitude of the degrees of freedom, exceeding the conventional standard for goodness of fit. With this fit, Equation 4 establishes a close link between the numbers and the data (although the model is not necessarily unique).

For the height-weight exemplar in Figure 3, Figure 6 shows the successive improvements of fit as the features of Equation 4 are implemented in stages.

***F* 'Rule'**	**Chi-Square**	**Standardized Chi-Square**	**Chi-Square/ DF**	**Number of Parameters**	**Degrees of Freedom**
Bi-Variate Normal Assumed: **Wt= 8.369 lbs/inch*Ht –394.912 pounds f1**	***879,401***	**32,357.62**	**2,383.20**	**6**	**368**
Best-Fit Null[1] f2	***1,127***	**30.94**	**3.40**	**44**	**330**
… as above plus Best-Fit attenuation, *a = 1.1* Estimated: **Wt= 4.290 lbs/inch*Ht – 157.394 pounds *F4***	***248***	**-3.45**	**.76**	**47**	**327**

Figure 6. Reduction of Error Corresponding to Features of the rules for height-weight data in one dimension)[2] - The parameters of the two linear transformations, m, n, p, and q, resolve

[1] As *every* model based solely on row and column multipliers is a null model, the best-fit null is the null model that best fits the data using the same criterion (least chi-square) that is used to assess the positive model.

[2] There is no probability calculation implied by these empirical chi-square values: Probability tests would require that the cell values be Poisson distributed with means greater than approximately 4 or 5 which is not the case in these data.

in to two parameters describing the linear relation between them and use only two degrees of freedom.

Falsifiable Hypotheses:

Transferring this non-Gaussian rule to a network for which numbers are yet to be found, the Washington Post's 911 data were prepared as a frequency table showing the number of links (activities) shared by each pair of individuals, Figure 7. The fitted frequencies are shown in Figure 8. Finally, for these frequencies the best-fit two-dimensional SCM displays the not-previously-estimated numbers, Figure 9. The SCM of not-previously-measured numbers is backed up by a close fit to the data, chi-square ≈ 1.66 — a close fit (To work in two dimensions, the distances on the line, for height and weight, were replaced by Minkowski distances in two dimensions, with coordinates yet-to-be-estimated[3])

It displays some subjectively familiar features: A dense 'clique' combining parts of the two Word Trade Center groups, a separate 'clique' for the Pentagon, and no structural coherence for the three-member group that failed.

Addressing the primary question, does the evidence support the hypothesis that the SCM with not-previously-measured *x's* and *y's* is an objective representation of this network? The evidence from the fit is consistent with that hypothesis: Given these estimates of the *x*'s and *y*'s, the hypothesis achieves a close fit to the frequency data in much the same way that an ordinary linear model, with known *x*'s and *y*'s, might achieve a close fit to the means.[4] For these data, the close fit is a strong argument in support of the attenuation and Minkowski parameters of the model (Appendix I), in support of the reality of the space, and in support of the estimates of these not previously estimated *x*'s and *y*'s.

---

[3] To generalize Hidden Line methods to 2 or more dimensions the absolute difference

$$d_{ij} = |x_i - y_j| \qquad (4)$$

is generalized to the Minkowski distance — there being no reason to assume that the geometry of Euclidean space (used for physical space) is a good geometry for other data spaces.

$$d_{ij} = \left(\sum_{\dim=k}^{ndim} |x_{ik} - y_{jk}|^M\right)^{1/M} \qquad (5)$$

This is the family of Minkowski metrics. Their different forms of combination, parameterized by *M* allow exploration of best-fit real-world rules of combination. If M were equal to 2, if it fits, the metric would imply that dimensions of a data space, like dimensions of physical space, combine according to the square root of the sum of their squares. If the best-fit *M* were equal to 1, it would imply that the dimensions of a data space combine by straight addition of their orthogonal components.

Observed Frequencies

-	(6)Hanj	(6)Moq	(6)Naw	(6)Sale	(7)Almi	(12)Atta	(3)Wale	(3)Wail	(3)Suqa	(3)Alom	(10)Al-S	(3)Bani	(3)Ahm	(4)Ham	(3)Moh	(3)Jarra	(3)Alna	(3)Saee
(6)Hanjour (	.	4	2	4	2	0	0	0	0	1	0	0	1	0	0	0	0	0
(6)Moqed (F	4		2	3	4	1	0	0	0	1	0	0	1	0	0	0	0	0
(6)Nawaf Al	2	2		2	5	0	0	0	0	0	0	0	0	0	0	0	0	0
(6)Salem Al	4	3	2		2	1	0	0	0	1	0	0	1	0	0	0	0	0
(7)Almihdha	2	4	5	2		1	0	0	0	0	0	0	0	0	0	0	0	0
(12)Atta (W	0	1	0	1	1		3	3	2	2	9	1	0	0	0	2	0	0
(3)Waleed A	0	0	0	0	0	3		3	2	1	3	1	0	0	0	1	0	0
(3)Wail Alsh	0	0	0	0	0	3	3		2	1	3	1	0	0	0	1	0	0
(3)Suqami (	0	0	0	0	0	2	2	2		1	3	2	0	0	1	1	0	0
(3)Alomari (	1	1	0	1	0	2	1	1	1		1	0	1	0	0	1	0	0
(10)Al-Shet	0	0	0	0	0	9	3	3	3	1		2	0	0	1	1	0	0
(3)Baniham	0	0	0	0	0	1	1	1	2	0	2		0	0	2	1	0	0
(3)Ahmed A	1	1	0	1	0	0	0	0	0	1	0	0		2	0	0	0	0
(4)Hamza A	0	0	0	0	0	0	0	0	0	0	0	0	2		1	0	1	1
(3)Mohand A	0	0	0	0	0	0	0	0	1	0	1	2	0	1		1	0	0
(3)Jarrah (F	0	0	0	0	0	2	1	1	1	1	1	1	0	0	1		0	0
(3)Alnami (F	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		3
(3)Saeed Al	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3	

Figure 7. Numbers of Activities Shared by each pair – Observed Frequencies

	(6)Hanj	(6)Moq	(6)Naw	(6)Sale	(7)Almi	(12)Atta	(3)Wale	(3)Wail	(3)Suqa	(3)Alom	(10)Al-S	(3)Bani	(3)Ahm	(4)Ham	(3)Moh	(3)Jarra	(3)Alna	(3)Saee
(6)Hanjour (		4.2	2	3.9	1.9	0	0	0	0	1	0	0	1.1	0	0	0	0	0
(6)Moqed (F	4.2		2	2.9	4	1.4	0	0	0	0.7	0	0	0.8	0	0	0	0	0
(6)Nawaf Al	2	2		2	4.9	0	0	0	0	0	0	0	0	0	0	0	0	0
(6)Salem Al	3.9	2.9	2		2.1	1	0	0	0	1.1	0	0	1	0	0	0	0	0
(7)Almihdha	1.9	4	4.9	2.1		0.7	0	0	0	0.6	0	0	0.1	0	0	0	0	0
(12)Atta (W	0	1.4	0	1	0.7		3	3	1.9	2.1	8.9	1	0	0	0	2	0	0
(3)Waleed A	0	0	0	0	0	3		2.7	2.2	0.9	3.1	1	0	0	0	1.1	0	0
(3)Wail Alsh	0	0	0	0	0	3	2.7		2.2	0.9	3.1	1	0	0	0	1.1	0	0
(3)Suqami (	0	0	0	0	0	1.9	2.2	2.2		1.1	2.6	1.9	0	0	1	0.9	0	0
(3)Alomari (	1	0.7	0	1.1	0.6	2.1	0.9	0.9	1.1		1.1	0	1	0	0	0.7	0	0
(10)Al-Shet	0	0	0	0	0	8.9	3.1	3.1	2.6	1.1		1.9	0	0	0.9	1.3	0	0
(3)Baniham	0	0	0	0	0	1	1	1	1.9	0	1.9		0	0	2.2	1	0	0
(3)Ahmed A	1.1	0.8	0	1	0.1	0	0	0	0	1	0	0		2	0	0	0	0
(4)Hamza A	0	0	0	0	0	0	0	0	0	0	0	0	2		1	0	1	1
(3)Mohand A	0	0	0	0	0	0	0	0	1	0	0.9	2.2	0	1		0.8	0	0
(3)Jarrah (F	0	0	0	0	0	2	1.1	1.1	0.9	0.7	1.3	1	0	0	0.8		0	0
(3)Alnami (F	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		3
(3)Saeed Al	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3	
1st Coordin	0.76	0.37	0.72	0.67	0.34	-0.25	-0.73	-0.73	-0.68	0.19	-0.64	-0.36	0.41	0.26	-0.64	-0.64	0.53	0.56
2nd Coordir	-0.12	-0.55	-1.29	-0.14	-0.58	-0.23	0.17	0.17	0.24	0.21	0.12	0.67	0.39	1.41	1.17	0.22	2.25	2.26
Multiplier	2.8	2.2	294797	1.4	1.8	5.2	1.6	1.6	1.4	1.2	1.9	1.6	0.9	4E+07	1.8	0.7	0	163
Error v/s null	0.000	4.407	1.133	7.658	16.748	7.456	6.044	10.378	7.904	1.990	19.252	7.754	6.669	18.549	14.941	7.370	9.452	58.901
Error v/s Mo	0.000	0.011	0.001	0.005	0.018	0.240	0.000	0.046	0.054	0.720	0.073	0.024	0.187	0.000	0.045	0.239	0.000	0.000
Chi-Square	1.66	Attenuation		12.64	Minkowski M		0.81											

Figure 8. Numbers of Activities Shared by Each Pair - Fitted Frequencies

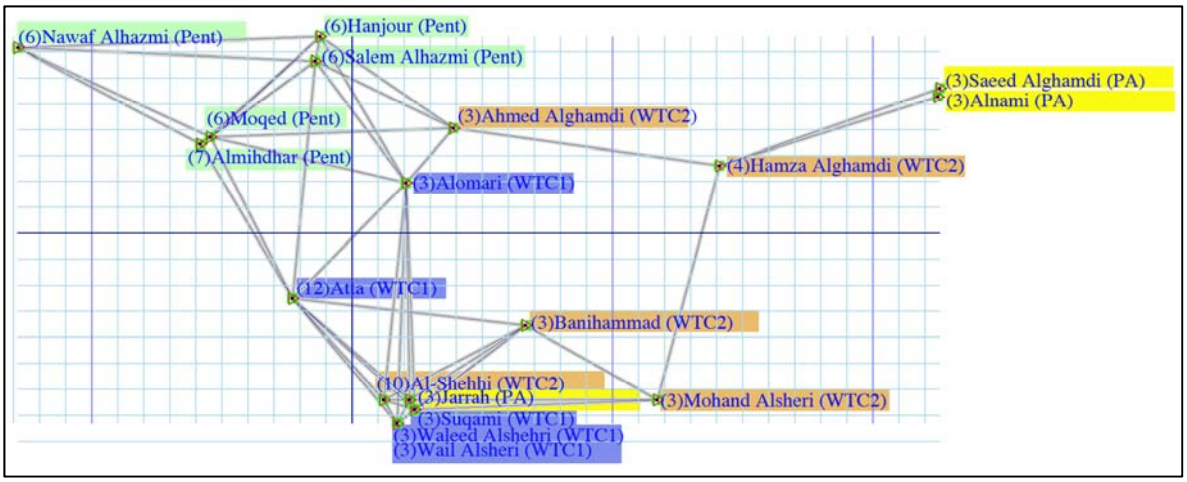

Figure 9, Map: Inferred Coordinates for 911 Terrorists - Attenuation 12.64, Minkowski parameter .81, Chi-Square = 1.66

Why Measure?

Putting numbers on not-yet-measured objects is not an end in itself. It is important, in part, because it joins measurement with theory. In turn that puts a data analysis in jeopardy, as it should be: "Methods" are not theory-neutral. Joining measurement to theory which means it can fail, which means it can be improved — which means it can extract simplicity (in the height-weight case) that the creative ambiguities of English can not detect. In contrast, the proposed SCM approach will lay bare the relations and help the analyst to develop empirically driven theory in which ambiguity is reduced through the systematic assessment of alternatives.

LINEAR MODEL VISUALIZATION IS OBJECTIVE

Current mapping technologies using standard graph theory and social network visual analytics do not support formal inference and deduction of facts not explicitly present in the data.. Current mapping techniques also have trouble simultaneously handling both networks and attributes, and then using the position of the nodes on those attributes both to infer networks and to interpret the results of findings about the position of actors in the network or the composition of subgroups. Finally, many social network visualization methods, e.g., force-field layouts, invite the analyst to change the visualization, in pursuit of visual clarity – which is subjective. The force field layout allows the user to alter "gravity" and "repulsion" to suit. This makes them subjective, by definition: The result depends on the observer.

To be sure there is a body of research aimed at two-mode data and methods exist for creating network from bi-partite graphs (e.g., multiplying a network by its transpose). And that resolves part of the visualization issue. More relevant to the work proposed here there are a number of techniques for creating distance networks. A distance network, is a matrix of relations among nodes such that the link weight represents the "distance" between those nodes given a set of indicators. Common distance metrics include similarity, relative

similarity, Euclidean, Chebyshev, Canberra, and Minkowski. The metrics vary in the extent to which they weight outliers, are valid for continuous versus categorical data, and control for correlation among variables. We note that there is no agreement on which metric to use. Thus, a key element of our research will be to identify the best candidates and assess the sensitivity of the SCM results to this metric.

The visualization results achieved from the use of linear models and the SCM approach are quite different than those typical in social network analysis. In ordinary geometry, if the distances between one object and three different objects are known, then the distance between that object and all different objects can be inferred, whether or not data are provided for these additional distances. Further, if data exist for the distances between one object and four or more different objects are known – and there are errors in the data, then the data can be corrected, because the correct values must be consist with the Euclidean rules. It is for this reason that satellite navigation systems use as many satellites as are available. In contrast, current "network geography" techniques which are used to assess topic-maps and social-networks do not support this type of formalized inference and deduction as there is no meaning associated with the position of the nodes in the 2D or 3D visual image.

If a semantic or social network is represented as shown in Figure 10, visualization1, which is a typical network visualization, this typical visualization is not true to all that we know: In this visualization, A, B, D, and E are equidistant from each other, through the center node - C. But the visualization does not show that: It shows A closer to B than to E. Further, C is perceived as critical as it is in the middle. Finally, any information about the strength of the relation or the basis of the relation is missing. In Figure 10, visualization 2 which is an enhanced traditional network visualization, color and weight are used to provide additional information. But again, the image is not accurate as the length of the lines connecting the nodes has no inherent meaning. One cannot infer that node A is twice as different from node C as is B. We propose to develop a SCM network visualization approach where the nodes are place in either 2D or 3D space, and such that the nearness of the nodes to each other reflects similarity in this space; where inference can be drawn based on position; and where missing data can be estimated given the model expressed in this topic space. A stylized interpretation of this is shown in Figure 10 visualization 3 which is the SCM visualization. Our proposed approach will generate models where formalized inference and correction for missing data is possible and distance meaningful.

Figure 10, visualization 3 is the result of the linear model visualization. This visualization, like the real data example shown in Figure 9 - is objective. These representations are 'objective' in the sense that once data decisions are made, e.g., once it is decided that the frequency table is to be mapped, decided that the diagonal will be ignored (ignoring the vents that a node share with itself) – decisions that the user can choose, the placement on the map is objectively determined. It adds a degree of objectivity to network analysis, an objectivity that has been lacking with network visualizations. In the 911 data, there might be some visual appeal to spread the tangled web of the five people at bottom center. But that decision is not ours and not made on subjective grounds.

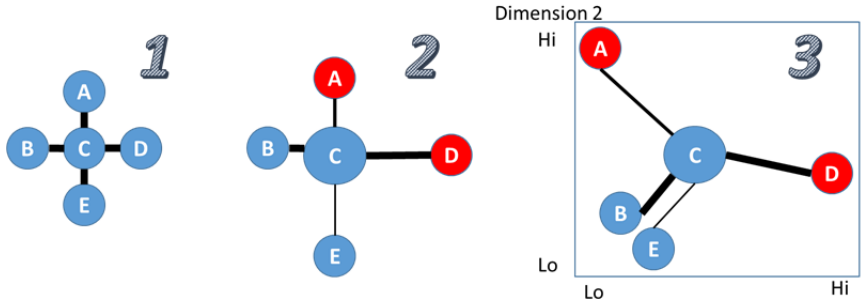

Figure 10. Illustration of different network visualization styles

In these SCM visualizations we are trying to display non-Euclidean images on a Euclidean piece of paper. Imagine the above images on a city-block grid. In city block metric, visualization 1 (in Figure 10) would have AB and D all equidistant from each other. Consider drawing a grid of 'city blocks" in which each was two units away from every one of the other three.

What does the SCM model tell us?

We can address that question in terms of both theory (the rules) and practice – the two come together in the attenuation parameter, "a".

Consider what it is not: a is not 2. In simple one variable distributions, "2" and the bell-shaped curves it describes, go with the informal interpretation of the Central Limit Theorem to the effect that when an observed variable is an aggregate (a sum or average of several independent and identically distributed contributing variables) it will approximate a "normal" distribution, a = 2. This thing is not normal, it has a 'spike' as it crosses the central line, where the "normal" would be flat. The suggestion is that weight is not an aggregate, and is relatively simple: With few (or highly correlated) contributing variables that are themselves "spiked". It suggests a research path looking for few causal variables. That is very different from the standard story of regression analysis: Use one variable, pick up a couple of percent of the variation. Add another, pick up a couple of percent — in years of further research, the "percent variance explained" will gradually improve. This spike is a different story.

Also note that the chi-square error dropped at a = 1.1 and, at the same time, the slope of the linear relation dropped from 8.5 lbs per inch (standard least squares regression) down to 4.3 lbs per inch, almost 50%. And note that one thing that was clearly wrong with the standard analysis is that it includes some women whose relatively heavy weight is not associated with height. Ordinary least squares and the normal do not "know" what to do with this: There is a something 'going on' in these data that has nothing to do with the relation between height and weight.

These extreme cases will affect the column means and, therefore, the least squares regression line. By contrast, these extreme cases do not affect cells associated with the

"spike" (from a ~ 1.1). The non-Gaussian 'rule' is associating the line with that part of the data that exhibits the spike.

Where the strategy is consistent with the data, what does the strategy teach about the world from which the data come? What do we learn that is not taught by the pre-computer statistical devices of classical data analysis? What does the strategy simplify and advance?

One thing it allows is a distinction between prediction and process: The least square best fit line is designed to predict averages. The linear relation that allows this model to fit the data describes descent on either side of a linear relation between x and y. But there is no reason to assume that these two lines are the same. For height and weight they are not. There is a ridge associated with this linear relation literally changes our description of the world, or offers a competing description of the world that generates the data.. Anticipating subsequent analyses, the "deepest" questions may involve that "a" parameter.

The "a" relates to the seldom explicit but often taught "story line" of behavioral research. The story has it that the world is a very complicated place wherein what we see is the result of many variables acting to produce the behavior we see. The story tells us to predict "normal" (bell-shaped) scatter among values surrounding predicted values — because, roughly speaking (very roughly) that is what the Central Limit Theorem tells us to expect from phenomena that are the aggregate of many underlying phenomena. The story line tells us that one generation of scholars will settle for explanations that explain some "percent of the variance", to be followed by the next generation "explaining" another couple of percent of the variance that is left, followed by …

Where a is not equal to 2, the story changes. Then the logic runs backward: If we are not seeing bell-shaped distributions, a = 2, then our phenomena are (or may be) simple. They may indeed be the result of many underlying variables, but those variables will be correlated such that the number of independent determinants of behavior is small. When a ≠ 2 fits the data it suggests, but does not prove, that "the world" is simpler than we have assumed, not simple but more so.

The extension of Equation 4 to multiple dimensions generalizes distance from the one-dimensional expression

$$d_{ij} = \left| x_i - y_j \right|$$

to the two or more dimensional expression

$$d_{ij} = \left(\sum_{k=1}^{ndim} \left| x_{ik} - y_{jk} \right|^M \right)^{1/M}$$

This is the family of Minkowski metrics for distance. They express properties of the underlying data, specifically the rule by which differences on several dimensions combine.

For M = 2, distance components of distance are combined by adding their squares and taking the square root of the sum — the Euclidean distance. For Euclidean distance each component contributes to the combination in proportion to its square so that larger components dominate the combination.

For M = 1 distance components combine by addition, the so-called Manhattan metric in reference to roads arranged in a rectangular grid. For Manhattan distance each component contributes to the combination in strict proportion to its size.

For 0 <= *M* <=1 distance components combine by adding their roots and putting the sum to a power. It can be referred to as the "can't get there from here" metric because the shortest distance between two points lies through an intermediary third point. (These metrics are properly referred to as semi-metrics as they relax the triangular inequality satisfied by ordinary (physical) distance.

Whichever metric applies, if one metric produces a better fit to the data, that fit reveals theoretical information about the underlying data. Examining combinations of attenuation and metric, the best combination for these 911 data is the Manhattan metric, M ~ 1.0, and attenuation slightly flatter than the normal attenuation, a ~ 3. (With a chi square of approximately 3, accumulated from153 cells, errors this small can only be approximate.) The impact of attenuation and the chosen metric value is shown in Figure 11.

	Attenuation				
Metric	***a* = .7**	***a* = 1**	***a* = 2**	***a* = 3**	***a* = 4**
***M* = .70**	87.82	22.80	15.62	3.96	4.07
***M* = 1.00**	122.99	8.06	3.14	2.85	1.99
***M* = 2.00**	108.46	8.29	4.24	4.25	5.31
***M* = 3.00**	110.56	6.80	4.89	3.42	4.55
***M* = 4.00**	92.12	6.05	4.21	6.54	What happened here?

Figure 11, Chi-Squares - Least Chi-square values corresponding to each of 25 fixed combinations of the metric and the attenuation.

MINKOWSKI PROCESS

The core of the SCM process relies on the Minkowski metric (or semi-metric). This section describes the underlying mathematics.

R_i = row i multiplier

C_j = column j multiplier

d_{Mij} = the Minkowski metric (or semi-metric), with Minkowski parameter M for distance from row i to column j

a = attenuation – this is the power to which the Mikowski metric is raised

$\hat{F}_{ij}$ = fitted frequency for cell ij

$x_{i,dim}$ = x is the row coordinate for row i

$y_{j,dim}$ = y is the column coordinate for column j

M = Minkowski parameter

Ndim = the number of dimensions – This will be 1, 2 or 3.

Chi-square = the sum over all relevant cells of (frequency – F_{ij})^2/ $\hat{F}ij$

Now the fitted frequency can be calculated as:

$$\hat{F}_{ij} = R_i C_j e^{-d_{Mij}^a}$$

The above equation is often referred to as the basic model. The alternative is to use 2 rather than e in the above function. Sometimes it is easier to get a distance of "1" on the SCM correspond to a half-distance decline of frequencies.

Finally, the Minkowski distance parameter is calculated as:

$$d_{Mij} = \left[\sum_{dim=1}^{Ndim} \left| (x_{i,dim} - y_{j,dim}) \right|^M \right]^{1/M}$$

AUTOMATING THE PROCESS

The implementation of these strategies is computationally intense — too costly for our predecessors, increasingly accessible to us. It leads to more ambitious appetites for what can be measured, to rules that can be stated, applied to data, and tested, and to a class of 'methods' that are half methods and half theory, able to extract more order from our data than was previously known to exist.

What are the nuts and bolts of implementing this approach to previously unmeasured behavior? This section provides the result of a cognitive walkthrough that will include, step by step instructions, to create an SCM.

In the SCM process the analyst will start by inputting a one mode network with attributes or a two mode network into the system, and then using the automated workflows which will direct the analyst through the creation, analysis and visualization and forecasting phases for the SCM (see Figure 11). Users with multiple data sources will enter two or more data matrices and will then be routed through the enhancement

workflow (see Figure 12. In Figure 11, SCM creation is a transformation and editing process by which the source data in what ever form it is in, is converted into a frequency table. Note that if the input is a binary actor x topic matrix (M) then this is created by multiplying M by its transpose, and then generally removing the diagonal. But as will be seen, depending on the nature of the data this may be more complex.

Figure 11. High level workflow for SCMs. SCM relevant technologies are in red. Note the assessment and visualization technologies mostly exist and are in ORA and will be re-used. This workflow hides, what we estimate to be about 50 low level tasks that are currently not automated, but that will be automated in the proposed system.

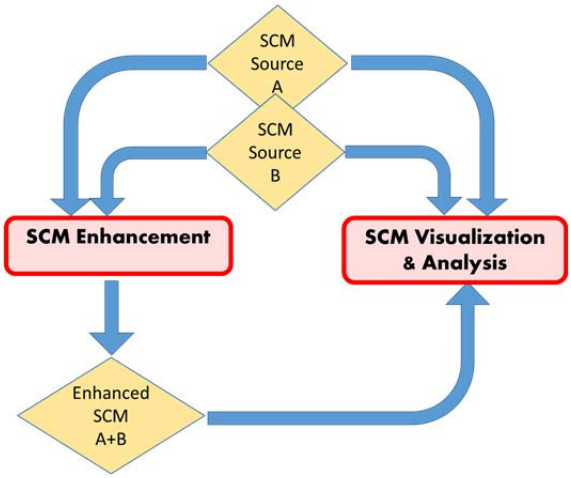

Figure 12. High level conceptualization of enhancement. The enhancement section of the process will be automatically called when the user enters or selects two or more actor x topic matrices. Specialized analysis and visualization tools will be used to support comparing and contrasting the original SCMs and the unified enhanced SCM. This will involve presenting the SCM visualization with annotations, creating a specialized report that provides information on the fit of the SCM to the raw data, key nodes, etc. And, this will involve the ability to not just spatially visualize the SCM but to visualize difference in two SCM's or overlay them.

Step 0: Keys in identifying attributes for creating an SCM from node by attribute data.

In this step the user chooses what they want to use as nodes and attributes. The nodes are the entities that will be displayed when the SCM is visualized. If this is actor x topics then the nodes might be actors. The attributes are the things that are, in a binary sense, true or not true of that node, e.g., whether or not they are concerned with that topic. When we are referring to the raw data we will use the term question, and reserve the term attribute for the final binary indicator.

For this process, the user starts with the raw data. We assume that the raw data ia already in ORA and that each question from a questionnaire or variable from a coding scheme is its own column. We assume that the nodes of interest are the rows. Note – the original raw data can be very messy and different solutions may be needed for each attribute in the raw data. For example, education may be a single attribute but it is categorical with the categories less than highschool, highschool, college, Ph.D. other. This could be converted to 5 binary attributes. In contrast, in the raw data identity might be

coded across many different attributes such as Suni, Sufi, Christian … and the node may have a score on each. In this case in the SCM creation process each of these original attributes might be saved as the binary attributes. If the raw data is originally binary: e.g. Column ?? are you Sunni. You can use it to produce 1 attribute (maybe 2 if non-response is a 2nd). And, of course, the raw data could be continuous.

The user has many choices here:

- How are non-responses or missing data treated? Should they become their own attribute? And if so, is there one such attribute per original question with missing data or non-response. Note – the default is to ignore this and treat non-response or missing data as a 0.
- Does the question have a binary response. In this case, it is already an attribute and used as is. Or, if there is missing data or non-response it can be converted into two attributes.
- Does the question have a categorical response. In this case each category becomes an attribute (default) or alternatively the user can choose to create an attribute that is 1 if the categorical response was greater than the mean or mode and 0 otherwise.
- Does the question have a continuous response. In this case a binary attribute is created by putting a 1 if the response is >= to the mean or else 0 (default), option 1 – allow the user to define categories of responses that form an attribute e.g. if < 18 then attribute youth = 1 else 0, if 19-30 then attribute young adult = 1 else 0, and so on.
- Are there meta-attributes created by combining answers to questions – e.g. <18 and sunni.

In general, we assume the user has an actor by attribute matrix. But it could be any node class with a set of attributes. We use actor by attribute to describe the process. The walkthrough revealed that it does not make sense to automate the process of defining attributes – but automation should provide some support tools. In ORA when there are node attributes we will let the user select a set of these and then have ORA create the "binary" file needed for SCM.

Case 1: the attributes are binary or can be converted to binary. If not, the user needs to convert them to binary.

To do this. First select the set of attributes. Then for each attribute do the following. If it is already binary but text convert to binary numeric. Let the user choose which string will be a 1 and which 0. If it is not binary and not text, tell the user they cannot use it. If it is binary and numeric use as is. If it is categorical, allow for three options: convert each category to its own attribute, let the user select a category, collapse the categories to binary using >= mean is 1 and else 0. If continuous then convert to binary by if value is >= mean then 1 else 0.

In an actor by attribute file – the cells are binary – i.e. the actor either has the attribute or they do not. Note missing data can be treated as an attribute. Make this an option for the user. In this step the SCM does not need to distinguish between missing data and 0's

Case 2: the attributes are not binary and are to be left as non-binary. This is an extension that will not be initially dealt with.

Step 1: SCM Binary input is selected

Get the actor by attribute file

For Syria this is (this is the identity question - 800 People by 17 Attributes (11 identities and 6 educational levels)

For 9-11 this is 18 people by 26 attributes

Case 1: Cells are binary - such as that shown in Figure 1.

Case 2: Cells are non-binary

We consider this an extension and will not handle it in V1. We will consider this in V2.

Then given a binary matrix use it to generate a frequency table. Note, at this point the user can enter the process by selecting a two mode network that is binary or can be converted to binary by setting each cell by if value is >= network mean then 1 else 0. In addition. At this point the matrix should be saved as a two mode network.

Step 2: Create the Frequency Table

- Input is binary matrix. This may be a two mode network that is binary (e.g., actor by knowledge) or a node set and the set of binary attributes from the previous step (e.g., actor by attribute). A third option is if you have a set of three factors e.g. actors response about education and actors responses about identity. In this case education might become the nodes and identity the other set of nodes.
- Output is a frequency table (e.g., actor by actor). In the case of option 3 just described the frequency table is two mode (e.g. education by identity).
- Or you can skip this step and use as frequencies a one or two mode network that already exists. Note - it is possible to use 0's and 1's in a binary one-mode network as if they were frequencies, where "1" might indicate a high (buy unmeasured) frequency of friendly behavior between two nodes.

Can't just do simple matrix multiplication

- If this were done then the diagonal would have to be converted to 0 but only for square matrices.

- You need to keep track of missing data for later analysis – use the ORA missing value number – something like -9.9999999999999999
- If A is the 800x17 matrix then AA' is not what is wanted.
- A'A is attributes by attributes
  - Identities x identities - In this case you want to just look at identities you would remove the rows and columns for education - probably would just start with 800x11 – shared identities
  - Education x education - If you just wanted education you remove rows and columns for identity – probably would just start with the 800x6 – if you zero diagonal it is empty
  - Relation of identity to education – this is 11x6 - this is the upper right of the 17 x 17

Step 3: Check to make sure the attributes are not mutually exclusive and if they are fix it

If the attributes are such that they are mutually exclusive then

- If all mutually exclusive it will generate missing data cells in the frequency – and the entire matrix is blank. Just code this as the the ORA missing value number – something like -.9999999999999999. Note that 0 can be a correct really value in this procedure, but negative values cannot show up.
- If only some are mutually exclusive this will generate a 0 cell that will distort the final map as it impacts the goodness of fit. All 0 cell's must be resolved. There are two approaches to resolving this:

1) don't create such categories. In this case tell the analyst which categories created the problem and see if they want to remove the category, or select another binarization approach, or add the categories.

2) if you need such a category the cell needs to be marked as missing data before mapping. Then mark the cell as missing – use -9.99999999999999999999

Result –a well formed frequency table. An example of which is shown in Figure 6.

If you are using as input a two mode network, instead of a nodeset by attribute matrix – you also want to make sure that no two columns are mutually exclusive. This is assuming that the rows are what you are forming the SCM on and you are treating the column nodes and links to those as attributes).

Note, in V2 you want ORA to store the choices on how the attributes were created so that the user can later go back and try a different approach. For example, it is not uncommon for the user to redefine which attributes to include – such as including or not including DRUZE in the Syrian data, or segmenting DRUZE into two different binary attributes based on whether they had high or low education.

Step 4: Check to make sure the frequency table is well formed and if not fix it

There are a bunch of checks here that are error check to make sure that the frequency file that is sent to the next step is well formed, the right size, etc. Note, if the frequency

table is generated via the ORA process – these checks should all pass easily. If the user is entering the SCM process at this point with a frequency table then, it may not pass. Also if the user starts and restarts the SCM process and messes up the ORA generated frequency file then it may not pass. A frequency file from ORA's perspective is just a weighted network. There are 3 types of frequency files that should be allowed:

- Square symmetric e.g. identity by identity – This is a one mode network.
- Square asymmetric e.g. a directed relation such as mobility occupation of father by occupation of son. This is two mode network.
- Rectangular – e.g. identity by education. This is a two mode network.

So first the system needs to identify what type of network it is and the size of each mode. That should be reported to the user and the information used to choose the path through the SCM procedure.

Second, the user should be asked whether the SCM process should try to fit the diagonal. In general, the default is that for a one mode network the diagonal is not fit and indeed it should be zeroed out and for a two mode network it is fit. The advanced option is to allow the user to choose.

At any point in going through the SCM workflow the user should be able to backup to the prior step or quit.

Step 5. Set parameters for optimization and the Minkowski procedure

Select Number of Dimensions

First, the user needs to specify the number of dimensions, Ndim, of the desired solution. At this point the options are 1, 2 and 3 and Ndim=2 is the default.

Select Function for Calculating Distance/Similarity

Second, the user needs to specify the metric for measuring distance and whether the SCM procedure can optimize or change that metric. If it is optimized or changed it is done with Minkowski. Note a set of similarity/distance metrics should be provided, and the default is to use the Minkowski approach. This is choosing the way in which similarity/distance will be calculated.

For distance, d_{Mij}, the default is to use the Minkowski metrics with parameter M.

For $0<M<1$, the name "metric" does not apply because the numbers for "distance" can violate the triangle inequality. In the help we will use the phrase "semi-distance" to name this thing. These semi-distances violate the triangle inequality but it is probably peculiarly appropriate to network analysis. Given paths diverging from a center, the shortest distance may be through the center, rather than directly across the coordinate's space. Therefore the indirect path, through a common center, may be shorter than a direct-looking path that tries to get between two nodes without going to the center.

In future we want to allow the user to select a set of metrics and then each is run in parallel. But -–this can cause a problem with local minimization.

Using the Minkowski distance is the default. Advanced options are:

Cauchy Option

Using the base model to calculate the frequency matrix is the default. An advanced option is to try instead the Cauchy form:

$$\hat{F}_{ij} = \frac{R_i C_j}{(1 + d_{Mij}^{a})}$$

This model worked spectacularly well on 4 examples from one of Green's books on correspondence analysis. Oddly, it hasn't worked well elsewhere, although I rarely try it.

This model is to the base model as "fat tailed" probability distributions are to the Gaussian. It is an option, but will rarely be used,

Other distance metrics

These are to be determined. These may not require a Minkowski parameter.

Select the Minkowski parameter – M

By default a Euclidean space is assumed and M is equal to 2. The user can choose to alter it. By default constrain M > 0, As an advanced option, allow the user to choose whether or not to enforce this constraint..

Determine the number of multipliers and coordinates

Third, the system needs to set how many multipliers and coordinates are needed so that the ensuing system will produce the correct number.

- For square symmetric the row multipliers and coordinates are the same as for the column. Thus is there are 4 rows/columns there are 4 multipliers and 4 coordinates.
- For square asymmetric the multipliers can be different but the coordinates are the same. In this case if there are 4 rows/columns you will have either 4 or 8 multipliers and 4 coordinates
  - You would want the multipliers to be different if there is a logical reason why the rows and columns are facing different issues. Otherwise you want them to be the same. So ask the user what is the case.
- For rectangular – the row multipliers and coordinates are different from the column's. So if you have 4 rows and 6 columns you want 10 multipliers and 10 coordinates.
- Note – how many "numbers" there are to a coordinate depends on the dimensions – so if there are 10 coordinates but 1 dimension there are 10 numbers, if 2 dimensions 10 pairs of numbers or 20 numbers, and if 3 dimensions 10 triplets of numbers or 30 numbers.

Select whether to fit the diagonals

If the diagonals are 0 do not fit, else fit them.

Step 6. System determines whether it will fit the diagonal

- If the frequency file is square symmetric
  - If the diagonal is 0 and then you don't try to fit it. Note this is the default.
  - If the diagonal is not 0 then the user should be given the choice to fit it or not; e.g. if diagonal is "different in kind" and not just the result of folding then don't try to fit it. The ORA SCM process will know this if it has been used to create the frequency file.
- If the frequency file is square asymmetric. Fit the diagonal if it is non 0 otherwise do not fit it.
- If the frequency file is rectangular always fit the diagonal

Step 7. Generate multipliers and coordinates

These are generated automatically by the SCM process. The user is not involved directly. This step basically sets the initial values as the optimizer will change them.

Step 8. Calculating the Chi-Square

In the next step, an optimization function is run to minimize the chi-square. This chi-square is based on a table of data. The data is the frequency table from step 2 that has been checked through steps 3 and 4. The "Observed" in the chi-square are the cells in the frequency table. The role of "Expected" values is taken by the values predicted by the SCM model.

$$\chi^2 = \sum_{relevant\ cells\ of\ data} \frac{(observed\ Frequency\ or\ frequency\ like\ datum - corresponding\ value\ from\ the\ model)^2}{corresponding\ value\ from\ the\ model}$$

Note – in later variants we might try things other than a Chi-square.

Step 9. Optimize the fit of the SCM

The input is the options just identified in steps 5 and 6, the frequency file, and the initial values for the multipliers and coordinates.

The goal is to generate a least-chi-square for the fitted value for the cells being examined. This does not satisfy the mathematical properties of chi-square and just using this as a convenience. So we should put a warning about that in the interface. Further, it is slightly difficult to calculate the degrees of freedom. For the initial tool we will not even try. For V2 we will include this calculation. With NR row multipliers and CR column

multipliers there are NR+NR – 1 of them that count against the degrees of freedom. Further. it actually matters whether or not the space is Euclidean with attenuation equal to 2. In this case fewer of the parameters are independent. See the section on determining the number of coordinates and multipliers.

The fit is a function of the multipliers, the coordinates, the Minkowski, the attenuation, given a model.

Optimization is used to select the multipliers, the coordinates (and maybe the Minkowski and the attenuation) that give the least-chi-squared. Optimization is a multi-step process. The goal of the optimization is to minimize the chi-square.

Why? Heuristically it works. Would a pure optimization approach be better? It is not clear as one needs to consider rate of convergence and so the speed of the overall system. The idea is to use this approach and then experiment with alternatives.

Heuristics Method: pick 4 possible values for Minkowski and 4 for attenuation – then for each of these 16 cells run the optimizer and find the multipliers and coordinates that minimize the chi-square

- The values to use as a default for Minkowski are .7, 1,2, 3, and infinity. As an option allow user to set their own values to try and there can be any number of these between 0 and infinity.
- The values to use as a default for attenuation are 1,2, infinity and .7. As an option allow user to set their own values to try and there can be any number of these between 0 and infinity.
- Good results often have attenuation being = to the Minkowski value minus 1.

Select the one these that led to the minimum and then run a second optimization where it changes all of the multipliers, the coordinates, the Minkowski and the attenuation. If the operation is fast to run, calculate all 16 cells, pick the best as the starting point and move out from there. If it is slow to run, start out only checking the cases: Minkowski 2 attenuation 1, Minkowski 1 attenuation .7 and see which is better – then move out from those in the direction that makes it better.

Possible optimizer to use is the simulated annealer. Note everything should be written in C++ like the rest of the tool.

Options for Optimizing the SCM

Log2 or ln

Use e as the default, and 2 as an option in the basic model. Note, this may not be the best default and alternatives should be explored. This issue is that if you use the same parameters you can compare absolute distances. If we've given them to different bases, there is going to be confusion. (If the distances are short, you get higher frequencies, even if those distances look exactly the same (proportional to) longer distances in a different map. A default of 2 is nice because regardless of the a parameter, a distance of 1 will give half the frequencies found at distance 0.

Multipliers

The multipliers should be set so that the geometric mean of the row multipliers is equal to the geometric mean of the column multipliers. Where the geometric mean of n multipliers is the nth root of the product of the multipliers. This is the default.

User can select this option. The option is to set the use an "All multiplier" applicable to 'all' cells, which would allow the row multipliers to be standardized to the geometric mean 1, ditto for the column multipliers.

Coordinates

When M is 2 you are in the Euclidean case. If M=2 then set the unweighted mean of the row coordinates to 0 (in each dimension). Same for the column coordinates. Note that, at M = 2, the effect that the row coordinates have on the model is invariant under an additive transformation. Ditto for column coordinates. Hence this standardization. At M = 2 for the row, only the intervals among rows matter. Ditto for columns. (Do not alter their scale, just their means.

If M is not equal to 2, then the coordinates in any one dimension, both row coordinates and column coordinates have to be treated together. In this case, subtract the unweighted mean of their coordinates from their coordinate – translating them together as a set. Do not alter their scale, just their joint mean.

Minkowski power

This helps to define frequency

- User selects
  - Starting value and allows the system to improve. As noted above the default is to use the 4 pre-defined values.
  - What functional model to use:
    - Do anything
    - Select the function from a list of those available

Frequency Attenuation

- This controls what power of the semi-distance is use for attenuation of the frequency
- User selects
  - Starting value and allows the system to improve. As noted above the default is to use the 4 pre-defined values.
  - What functional model to use:
    - Do anything
    - Select the function from a list of those available

Functional models

These are used as the options for the Minkowski power and the frequency attenuation.

The functional models to make available are:

- E to the attenuated distance
- Inverse power law
- Correspondence analysis
  - (Side note – use the correspondence analysis already written in ORA)
- Cauchy
- Others will be added in the future

Additional Option

Allow the user to turn off the attempt to calculate Minkowski power or frequency attenuation. By default this calculation is turned on.

Allow the user to let the Minkowski generate a 0 distance. By default a 0 distance is not allowed. Note – we may need to set this differently depending on the functional model. The distance can be controlled by not letting the optimizer set certain values or by using starting points in conjunction with an optimizer that can never reach 0.

Note a help file should contain the information in these steps but as explanations.

Calculate Error

This is the Chi-squared error. If the frequency table is square symmetric then calculate error on only the upper triangle. Else calculate error on the entire table.

Optimization Routine

It is very likely for this system to regardless of the optimizer chosen not settle into a single final value. In the end we may want to allow for multiple optimization approaches. However, in phase 1 - rather than annealing try this simple heuristic. It is like an annealer but without the cost function and the re-starts. Just simple hill climbing.

For each of 1, 2 and 3 dimensions - Set three initial values for the row and column coordinates, row and column multipliers based on the constraints in the document and for other variables. Note that there can be lots of parameters (including each coordinate). So if you have P parameters, “all combinations”, is going to be 3^P which could quickly be very large. Scalability needs to be checked for V2. If you are going to do a full evaluation of the table for each combination, the evaluation is the time burner, so we should check options in parallel

Run all combinations.

Rank order the results in terms of the fit of the chi-square. Find the set of these parameters that gives the best fit for that dimension.

Now going through the variables in this order - multipliers then coordinates slightly raise and lower the value while holding the other parameters constant. Within this "snowball" set - take the new value if the fit is better than the original. Continue to quiescence or 10 steps whichever is quicker.

Make it possible for the user to view the plots for how the fit changed as one or two parameters of interest changed.

As a side note on optimization, if there is no parallelization you can do this one parameter at a time as each parameter may impact only a small number of cells in the SCM. In contrast, downhill simplex (with or without annealing) is always a full evaluation simultaneously hanging all parameters. It has different costs.

Step 10: Visualization

Once chi-square is minimized, then you produce a visualization. This is done using the ORA visualization tool for grid-based visualization. The minimum chi-square is considered the best fit. An example of the visualization is shown in Figure 7.

Add a report that provides:

a) The values for all the Minkowski parameters.
b) The amount of error
c) Attempt to calculate the degrees of freedom and print that
d) Calculate the standardized chi-square – this is a convention. Ideally the chi-square is equal to the degree of freedom. Print the standardized chi-square
e) Do a t-test to compare the square root of two of degrees of freedom and the chi-square. Is it near 2? If so print out that this is a good fit.

EXTENSIONS

How far does this strategy go toward introducing numerical variables and testable hypotheses?

Beyond height-weight, a pedagogical "workhorse" of the statistics trade, and beyond this terrorist network, how far does this strategy take us into solutions for not-yet-measured variables? We suggest that this approach will have value in a wide number of areas, such as:

- to a re-conceptualism of sociological/political surveys as networks
- to pharmacology where it detects a dimension within the relation between drugs (functional groups) and biological effect
- to an examination of social structure and stability within the Syrian Opposition,
- to network analysis where it provides an objective goodness of fit with which to evaluate competing 'visualizations' of a network,
- to text analysis, and the analysis of real world budgets.