Applying Modality and Equivalence Concepts to Pattern-Finding in Social Process-Produced Data

Robert A. Hanneman
Department of Sociology
University of California, Riverside

Paper presented at the workshop on "Social Science and Social Computing: Next Steps" Honolulu, Hawaii, May 22-23, 2010.

Abstract

Large amounts of detailed transactional information are generated by ongoing social processes. Managing and mining such data treat them as “objects” and “relations.” These ideas strongly parallel the way that social network analysts conceive of social structure. Modality (roughly, distinguishing multiple classes of social actors or nodes in networks), and equivalence classes (roughly, distinguishing general patterns in the ways that objects in classes are related to one another or to objects in other classes) have proven to be very useful in helping social network analysts to think about complex relational structures among social objects. Dimensional and generalized “block models” of multi-modal social networks provide tools for designing searches to identify patterns. The ideas are illustrated by descriptions of how a number of social process produced data might be approached (e.g. Medline, game logs, relational data bases of transactions and summarized transactions).

Introduction

Every minute of every day, huge amounts of data generated by on-going social interactions are deposited in digital databases. These records are remarkable collections of "trace evidence" (Webb, et al., 1966) produced by social processes for their own purposes. While social scientists have always "mined" archives of records (e.g. manuscript censuses, newspapers, roll-calls of votes, mortality registers) as "non-reactive" ways of understanding patterns of social structure, the current era is unique in the amount of all social transactions that are documented, the accuracy of these records, and the sheer volume of data. Not surprisingly, the "mining" of digital archives and transaction logs is a very rapidly growing enterprise within and without the social sciences (e.g. new journals such as *Social Network Analysis and Mining*).

There are remarkable similarities between the ways that some social scientists think about the information in such archives as "pictures" of social structure, and the languages and logics to the computer scientists, database engineers, and others who have designed and built them. To date, however, communication between these two groups has been fairly limited. Most social scientists speak the languages of information sciences badly, if at all. The arcane languages and conceptual schema of the social sciences may be both unfamiliar, and seemingly irrelevant, to engineers. Often the goals of practically-oriented data miners (e.g. search, optimizing processes, assessing reliability) are quite different from those of social scientists (e.g. finding regular patterns and abstracting generalizations).

At one point the gulf between the two cultures is not so wide. On the computer science/engineering side, "social computing" seeks to build architectures consistent to support social transactions – and (usually implicitly) use theories of social structures. From the social science side, the field of social network analysis (and particularly network dynamics and agent-based modeling) have extensive experience in formally modeling and analyzing the kinds of data that are being produced by social computing – but little experience in exploiting the flood of data that have become available.

In the text below, we are going to look at one small part of how social sciences (particularly social network analysis) and social computing might inform one-another. First, we will look at a very concrete example from the two perspectives. Next, some strong parallels between critical concepts of data structures and social networker's conceptual schemes are discussed. The ways that social network analysts look at social computing data, and what they want to know from it are, in some instances, quite similar to some of the goals of data-mining. Two particular ideas from network analysis are then explored: modality (roughly, thinking about heterogeneous classes of social objects and their relations) and equivalence (roughly, what we mean when we say that two objects are similar to one another in terms of their relational patterns). Following this, we explore some examples of how the concepts might be (or, in a few cases have been) applied to mining social process produced data.

Bibliographic Data Mining and the Evolution of Scientific Communities

Relational data bases of periodical literature are now a critical part of the infrastructure of doing research work. For sociologists, bases like *Sociological Abstracts*, and *Web of Science*, are everyday tools of the trade. To the information scientist, the key issues are entry, storage, search, and reporting architectures and algorithms. To the social scientist, the database is an archive of trace evidence deposited by social actors in the process of producing "knowledge."

As a data object, a periodical literature database could be organized as a single table ("flat file") with a row for each new article that appears. Each row might contain a number of fields (e.g. first author, second author, journal-volume-number-pages, keywords, abstract, text body, and references. One could mine the database by specifying unions and intersections of sets of values on multiple attributes of the records to produce lists.

That description would make most database designers wince – it is an inefficient database architecture that would make it difficult and slow to extract useful information. However, the "traces" left by many very important social processes are recorded in essentially this form of cumulating lists of transactions as they occur. E-mail logs, lists of searches conducted by visitors to Amazon, contributions to blogs or to virtual communities (multi-user games, open-source programming communities), sales records, and stock trades are some examples. These "data structures" are very much like the marriage registers, birth and death records, crime reports, voting roll-calls, and other documentary archives that have been mined by social scientists. Many other very important data collections about social processes are simply aggregated transactions – annual tables of trade flows of commodities among nations.

To make mining more efficient, databases of periodical literature actually use object-oriented and relational concepts for their organization. Rather than a single table of transactions, each with many attributes, the data are organized in multiple tables, and the tables are connected by indexing attributes. One might have a table of authors, one of journal titles, one of articles (which might contain the abstract, body, and references), one of key-words. Individual authors might be linked to other authors (co-authorship), to one or more articles (authorship), which appeared in one or more journals at a particular time, with various combinations of key-words. For most bibliographic databases, articles are also indexed to other articles by way of cited-author or cited-article relations. This is the familiar relational database containing multiple indexed tables with a variety of one-to-many, many-to-one, and many-to-many relations between objects in the various tables.

Bibliographic data miners exploit the relationality of the objects in the data tables in a number of ways. A few examples suffice: the extent to which the articles published in one journal (over some period of time) cite articles published in other journals form a network of directed journal-to-journal citation ties. Eigenvector centrality of journals in this network is called the "impact factor" of a journal – and is critical to its desirability, its value as social capital in the career attainment of scientists, its advertising rates, etc. We may trace co-authorship patterns (author-author networks that count the number of times authors have written together), co-citation (the number of times that one author cites another in their articles), the prominence of particular authors, which articles cite which others (to find critical paths and key contributions in the development of discourse), and so on. One particularly clever application of this type is recent work by Chomei Chen (2006) that identifies "research fronts" based on bursts (and other factors) in two-mode article/key-term networks.

Now let's take the rather different perspective of a social scientist studying the development of science, who is seeking to exploit these data. To the social scientist (e.g. Collins, 1998) the information in the database are "trace evidence" of an ongoing social process that produced the data. As a "thick description" or narrative analysis, the analyst sees a complex process of co-evolution involving heterogeneous social agents interacting to "construct" social reality (very sorry if that phrase causes immediate headaches).

Roughly, the process looks something like this. Individual scientists (each of whom has a history), become interested in topics, interact with other scientists (at the same workplace, in professional associations) are influenced by the published work of others, and create a new article. They may form direct ties of working together to produce one or more articles, or work together indirectly (by citing one another). The new items produced cite previous articles, and so on. At the same time, journal editors shape the process by seeking high-impact contributions; themes and research problems evolve through combination and division. In short, it is a complex process of co-evolution in which scientists, specific articles, research problem areas, venues of publication, institutions where work occurs all shape the connections of the "web of science" as it changes over time.

Traditional "history of science" treats the process as an unfolding narrative of individuals, events, places, and texts co-determining and influencing one another. Social science approaches to the same types of data attempt to find patterns and commonalities in repeated similar causal chains by identifying types of individuals, events, places, and texts that frequently co-evolve in similar ways.

The perspective of the information scientist and the social scientist in looking at the same bibliographic data would seem to be very different. But there are some fundamental ideas in common.

Shared Concepts

Many sociologists (particularly social network analysts) might describe their perspectives as "object" or "agent" oriented, and focused on "relational structures." Perhaps somewhat surprisingly, they appreciate these terms in the same general way as computer scientists – though both groups have elaborations of the basic ideas that go in somewhat different directions.

For the sociologist, the "particles" that make up the relational structures they study can easily be seen as objects in very much the same sense intended by object oriented programming:

> "Object-oriented programming (OOP) is a programming paradigm that uses "objects" – data structures consisting of datafields and methods together with their interactions – to design applications and computer programs. Programming techniques may include features such as data abstraction, encapsulation, modularity, polymorphism, and inheritance." (Wikipedia, 2010)

The most obvious kind of a "social object," of course, is an individual human being. Persons have social identities described by attributes. Persons also have what social scientists are wont to call "agency," which is strongly analogous to the OOP notion of "methods." That is, persons have capacities to initiate behavior – and particularly behavior that creates, modifies, or deletes relations to other objects in the class(person), and other classes. The other intriguing concepts of OOP (abstraction, inheritance, etc.) don't obviously apply (but this would be interesting to explore!).

When thinking systematically about social structure as composed of objects and relations, sociologists (variably) also recognize some classes of "social" objects that are not people. Rather uncontroversial are the notions that "events" and "organizations" are social things with attributes and agency. "Events" are interactions that have their own emergent attributes are recognized by the actors (named, having shared meanings); for example, a research article might be thought of as an "event." The article has attributes (length, topic, co-authors, citations, etc.), a name in itself, and a "social life" of it's own that is not reducible to the attributes of the agent (s) that produced it. "Organizations" (couples, families, small informal groups, large formal organizations, whole nations, etc.) are also recognized as socially meaningful and have attributes and methods that are unique to their class.

More controversial, but regarded by many sociologists as very useful, is the idea of treating cultural objects as social objects. Identities, categories, and symbols (e.g. "engineer" or "American flag") are shared meanings that have attributes and emergent "methods" (this last is the controversial part, theoretically).

Sociologists often name what they study "social structure" or "patterns of social relations." Again, there is a strong analogy between the social science use of "relations" and the sense of the term when it is used to describe databases as structures of objects connected by indexing attributes or methods. Social objects (i.e. people, events, organizations, identities) are classes, and the patterns of relations among elements of a class, or between elements of different classes, are "social structure." The most explicit statement of this view of social structure is in social network analysis – where a social network is the set of social actors and relations connecting them.

The complexity of the social sciences lies primarily in the kinds of relations that are seen as connecting social actors. There is certainly no consensus within or between social sciences on classifying types of social relations. Social network analysis identifies two very abstract classes of relations: directed and "bonded." Directed relations or ties between two social actors indicate the conserved flow of some quantity from one to the other. A husband may direct money to a wife (and/or vice versa). "Bonded" relations or ties between two actors indicate that both are equally embedded in an "emergent social fact." A husband and a wife share the relation of "married."

There is a great deal that social scientists, and particularly social network analysts could learn from serious conversations with information scientists about the nature of "objects" and "relational data structures." But the two fields do have a great deal in common at a very basic level. Both are working with "structures" that are composed of "relations" (which have attributes) among "objects" (which have attributes).

The design and mining of relational data structures that are used to capture transactions of social processes is often approached by information scientists without thinking explicitly about the "social structures" that are producing the data. Social scientists think quite a lot about the processes of social structures that produce "data," but often lack the skills and/or motivation to exploit the data. In the sections that follow, I will suggest some particularly important organizing concepts from social network analysis can try to bring the two sides closer together.

A Social Network Analysis Approach to Relational Object Data Structures

The social networks perspective sees "social structure" as patterns of relations among social actors. These patterns are represented as graphs or directed graphs with nodes as social actors (who may have "color" spectra representing their attributes) and edges or arcs representing relations. Formal graphs have unambiguous translations into matrix representations. The "mining" or analysis of social network data consists of operations on these matrices to identify features of the graphs that are of theoretical interest, such as the "centrality" of nodes and graph "centralization" or partitions of nodes into classes based on similarities in their relational structures.

The notions of "modes" in social network analysis, and the kinds of relations they imply, are the basic conceptual tools that social network analysts use to think about how to organize complex relational data structures. There are many and varied tools for summarizing the patterns in the data (e.g. Wasserman and Faust, 1994; Hanneman and Riddle, 2005; Scott, 1991). For current purposes, we are going to focus on the problem of identifying (or testing hypotheses about) partitions of the data based on relational equivalence of social actors.

Representing Social-Process Produced Data Structures: Modality and Kinds of Relations

A large part of social network analysis focuses on the very simple data structure of a single relation connecting all elements of a class of social agents to other members of the same class. One can imagine a matrix of scientists by scientists, with elements containing the count of the number or articles on which they were co-authors. Structures that connect elements in a class to elements in the same class are labeled "one-mode" structures. In our example, scientists could be connected to scientists (in multiple relations such as "friendship" "co-authorship" "co-citers" "located at the same institution"). Articles could be connected to articles in one or more single-mode relations (one article cites another, two articles share authors, two articles appear in the same journal, etc.) Similarly, other classes of social actors could be connected in single mode relations (institutions to institutions, journals to journals, etc.).

Another data structure maps (one or more) relations between social agents of different types. The "two-mode" structure (e.g. scientists by articles, mapping who authored which) is rectangular. Two-mode data structures are also frequently called "co-occurrence" or "actor-event" or "affiliation" matrices. For some examples: authors are located at particular institutions; articles appear in particular journals; articles contain particular keywords.

Finally, a third common type of data structure, an "attribute" matrix, maps variables or attributes to the social agents in a class – giving the nodes "color". We might show the relation between scientists and the attributes of gender, ethnicity, numbers of prior publications, institution of employment, or whatever. In a multi-modal social network, there could be a separate attribute matrix for each mode (scientists have attributes, journals have attributes, institutions have attributes, articles have attributes, etc.).

We need to make a short side trip at this point, to talk a bit more about "attributes." Attributes of social actors, for example, the gender of a scientist, could be represented either as a "coloring" of the nodes, or as an "affiliation."

Table 1. Representation of Node "Color" as an Attribute and as an Affiliation

Person	Gender
Fred	1
Sylvia	0
Jonas	1
Jae-li	1

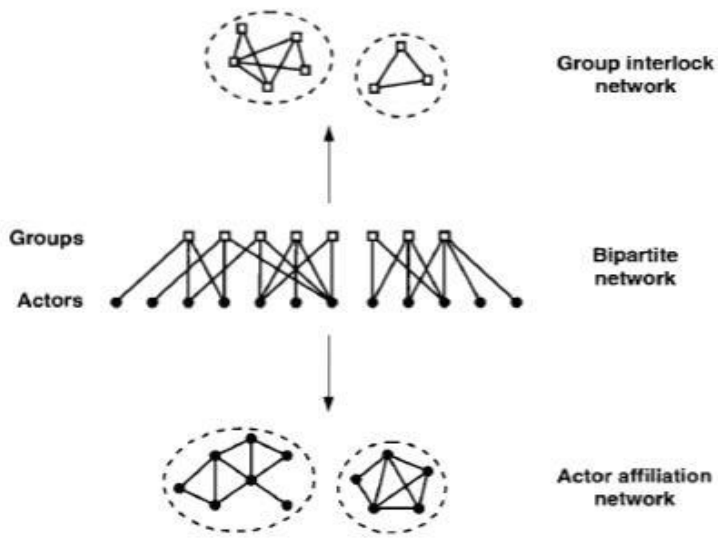
Person	Male	Female
Fred	1	0
Sylvia	0	1
Jonas	1	0
Jae-li	1	0

It can be argued that "color" should always be represented as affiliation, rather than as an "attribute". A person's gender, for example, is really an "affiliation" of a person with a cultural category or symbolic object – not something that is unique and wholly nested within that individual. The transformation between the two matrices above is trivial algorithmically. As a practical matter, it is often more insightful and useful to "color" nodes by attributes and use attributes as partitions. At a deeper level, though, many (most, all?) attributes of actors could be mapped as affiliations of actors with identities or cultural categories. When the goal of analysis is to find equivalence classes, as discussed below, it is often better to treat "attributes" of nodes as "affiliations" between two modes.

Relations in a single-mode matrix may be symmetric (represented as a simple graph with edges), or asymmetric (represented as a directed graph with arc). For example, the count of co-authorships between pairs of scientists is necessarily symmetric; the citation of articles by articles is necessarily asymmetric (though there may be reciprocal co-citation). Social action, however, is initiated by an individual and directed toward another. Thinking about social process suggests that one-mode social relations are best seen as directed and asymmetric. Symmetric relations among the elements of a mode of social actors can almost always be seen as induced from an affiliation matrix. For example, co-authorship ties between scientists might be though of as induced by affiliation of each scientist with the same object in another mode (the article class).

Thinking about social structures as mappings of relations among multiple modes, or different "kinds" of social actors, is the social network approach to dealing with the multi-level and qualitative complexity of social data. Ron Breiger (1974) provided the clearest and most compelling statement of the approach as a "duality" of persons and groups. Duncan Watts (a complexity scientist who recently migrated into social network analysis) represents the idea graphically.

Figure 1. Social Networks Approach to Modes and Affiliation



(Source: Watts, 2003)

Watts diagram illustrates that “groups” (which can be thought of as “events,” “organizations,” or “identities”) can be defined as having a relational structure as a result of the overlap of the agents that are affiliated with each of them. Actors also have an induced symmetric network by way of co-affiliation with the same events, organizations, or identities. Not shown in Watts’ diagram is the possibility that “actors” may direct ties to one another, and that “groups” may also have the “method” to direct ties to one another.

When a social network analyst is approaching social process produced data, if they were strictly following the logic outlined here, they would analyze the problem and create a straight-forward data structure:

- Identify the modes (qualitatively different types of social agents)
- Examine the “attributes” of each class of agent and treat them either as “colors” or as new “modes”
- For each mode, define a matrix of actors by actors by one or more relations. The relations may represent directed ties from one agent directed to another. If “colors” or agent “attributes” are treated as attributes (rather than as new modes) the resulting arrays will be symmetric. These arrays are square, but are simply a special case of the more general rectangular structure. Indeed, it is often very useful to treat actor-actor directed ties as two-mode matrices (rows as the mode “sender of tie” and column as the mode “receiver of tie”).
- For each pair of modes, define a matrix of the elements of one mode by the elements of the other by one or more relations. This will be a rectangular, asymmetric array.

These steps need to be understood as a proposal for defining data arrays to represent social structures, rather than the single, consensual, and "correct" way to translate problems. In reality, social network analysts are very ad hoc, and flexibly design their data structures to answer quite specific questions. Still, they might do well to be more systematic, because even quite complicated social processes can be reduced to comparable and understandable data structures by following the guidelines above.

The data produced by social processes then can be represented as some number of rectangular arrays of directed relations between the elements of each mode, and between the elements of each pair of modes. The arrays are linked by the indexes of the elements of each mode. Usually social network analysts will choose to retain some "attributes" of some or all of the modes – treating them as partitions, rather than relations.

Having structured the information, what data do we want to extract from it?

Mining Social Process Produced Data: Equivalence

In querying a database, we are locating data objects that satisfy (or are similar to) as set of criteria. "Show me all the books by Joseph Conrad, and are currently in print in paperback." It is easy to see such a query as asking about the attributions of a single mode of objects (books, in this case).

If we think about databases as relational structures or networks, however, the query might be understood a bit differently: "show me all book objects that have the relation "written by" to objects in the class "authors" with the attribute "Joseph Conrad," AND have the relation "true" to the object in the class "publication statuses" with the value "in print." We might imagine a three-way data array of authors by books by publication statuses, and ask to see the index values of all columns in the "books" dimension for the "row" "Joseph Conrad" in the author dimension AND the row "in print" in the publication status slice (that is, a specific value in the mode author; a specific value in the mode publication status; and any non-zero value in the mode book).

Making sense of complex relational data left by social processes can be seen as finding objects that are similar to one some prior hypothesis about relational equivalence (in a confirmatory analysis) or similar to one another (in an exploratory analysis). The book "Lord Jim" and the book "Nostromo" are "similar," in relational terms, because they are elements of the mode "book" that have an "authored by" tie to the element "Joseph Conrad" in the mode "authors."

But, what do we mean by "similar?" Social network analysts have given a good deal of thought to what it means for two social actors to be "similar" or "equivalent" in relational terms (Everett, 1994). Here, we will focus on the two most widely used definitions of relational similarity: structural and regular equivalence.

Structural equivalence was first explicitly define by Lorrain and White (1971) and is described in Batagelj et al. (2004) as: "Units are structurally equivalent if they are connected to the rest of the network in identical ways." Put even more simply: two nodes are structurally equivalent if they have exactly the same pattern of ties to all other nodes. Structural equivalence is the strongest form of equivalence –

exact equality in the pattern of relational ties. In practice, approximate structural equivalence is often used. There are numerous commonly used measures of approximate structural equivalence: correlation, Hamming distance, Euclidean distance, etc.

Almost all queries and methods of pattern finding (components analysis, cluster analysis, MDS, correspondence analysis) use some algorithm to locate dimensions, clusters, or classes of structurally equivalent nodes in graphs. In doing so, we are locating “substitutable” or “identical” nodes on the basis of their patterns of ties with other nodes. Almost all data mining, whether based on relational or attribute approaches, has used structural equivalence. Despite this, regular equivalence may be a more useful definition of relational similarity in many cases.

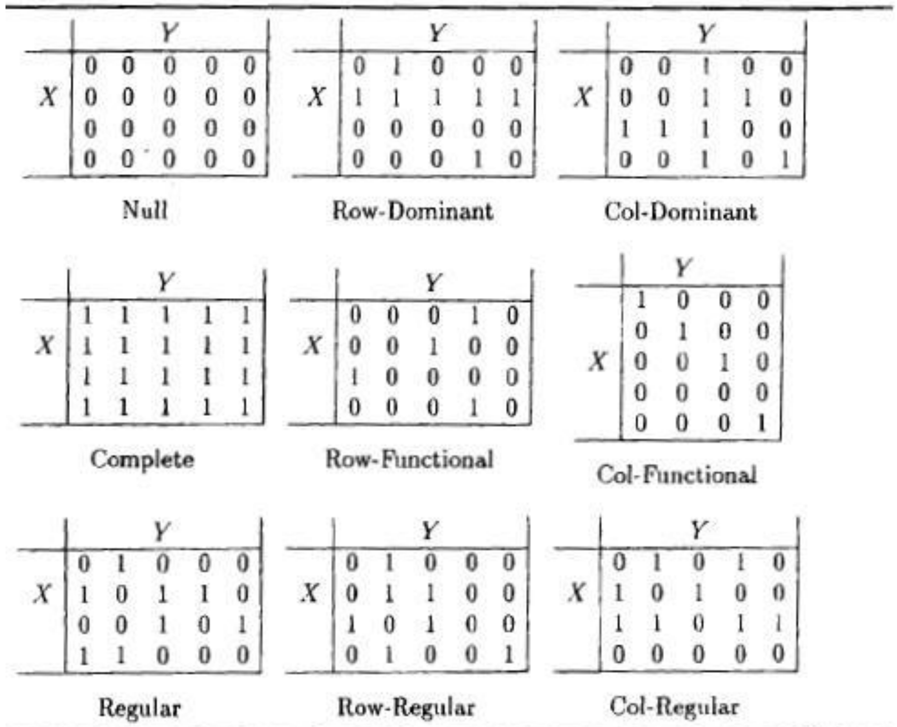
The first formal statement of relational regular equivalence is usually attributed to White and Reitz (1983). Regular equivalence, described in Batagelj, et al. (2004) as “…two units are regularly equivalent if they are equally connected to equivalent others.” The core idea is also sometimes understood with regard to the mathematics of coloring graphs. In graph coloring (Chung, 1997), two nodes in a graph are regularly equivalent (have the same color) if they have the same spectra (have at least one relation with an element of each the same set of other classes).

In social network theory, the idea of regular equivalence is tied to the notion of a social role. Consider a table that shows a list of adult women as rows, and minor children as columns. A cell contains a 1 if a particular child is the offspring of a particular parent, and zero otherwise. Using structural equivalence, no reduction of the rows is possible, as each mother has a unique set of specific children; reduction of the columns is possible, however, by grouping together the multiple children of a particular mother. Viewing the same data from the perspective of regular equivalence produces a different result. In this case, the adult women may be partitioned into two groups – those who have children, and those who do not; the minor children cannot be partitioned: each child has a relational tie to a member to the class of adult women who have children, and none has any tie to any of the adult women without children.

Regular equivalence is a “more relaxed” idea of similarity between nodes than is structural equivalence. In many cases, the goal of pattern finding and data mining is actually to find partitions that are regularly equivalent, not structurally equivalent. Regular equivalence is used to identify classes of actors who have similar “roles.” That is, they have similar patterns of ties to similar others. When we identify words or phrases as “equivalent” in the coding dictionary of content analysis, we are using regular equivalence; when we identify nations as “semi-peripheral” in the world system, we are using regular equivalence. Most social science theory is stated in terms of actors who regularly equivalent (e.g. “elite,” “parent”).

Algorithms and methods for testing hypotheses, or identifying regularly equivalent partitions in relational data are not as highly developed as those of structural equivalence. Probably the most commonly used approach is “block modeling.” In block modeling, the rows, columns, and slices of multi-modal graphs are permuted to locate blocks of cells that contain particular patterns of ties. One very useful example of the major types of blocks (or types of equivalence) is given by Doreian et al. (1994).

Figure 2. Relational Blocks in Generalized Block Modeling.



Source: Doreian, et al. 1994: 6.

The power of generalized block modeling in two modes can be illustrated rather simply. In the "core-periphery" view of economic relations in the world system, "core" nations export heavily to all other core nations. This would be a "complete" block of ties. "Peripheral" nations do not export to one another. This would be a "null" block of ties. Core nations each export to a sub-set of peripheral nations that fall within its sphere of influence, but not to all peripheral nations, generating a regular equivalence block. Peripheral nations export to some, but not all core nations, generating another regular equivalence block.

Social Network Analysis of Multi-mode Relational Object Data

The information produced by social processes can structured into multi-mode relational data. In these data structures, the goals of mining, generally, are identifying sets of cases in each mode that are equivalent (in either the structural or regular sense) with respect to the cases in each other mode.

Until fairly recently, social network analysts usually worked with multi-mode data by analyzing it one mode at a time. There can be great power in this approach.

Suppose that we were "mining" a data of email messages, and examining only the two modes of "sender" and "receiver." A rectangular array of senders and receivers is constructed (which would contain many, but not necessarily all of the same agents), and the presence/absence or number of messages in each dyad would be constructed. We could induce a matrix of which senders were similar to which other senders by counting the number (or volumes) of messages they sent to the same receivers. We could also induce a matrix of similarities among the senders by indexing the extent to which they received messages (or message volumes) from the same senders. Each of these "one-mode" square arrays could be thought of as a bonded (simple, un-directed) graph. Conventional network techniques could be used to identify central actors, and graph sub-structures (e.g. the "modular" community approach of Newman, 2006). Senders or receivers could be classified into groups or clusters based on similarities in the specific others to whom they directed messages, or to which other "types" of senders (or receivers) they were tied to. That is, the senders can be classified into either positions (structurally equivalent nodes) or roles (regularly equivalent nodes).

A great deal of interesting and useful information can be extracted by transforming the relational data for all pairs of modes into single-mode similarities. We can find senders who are similar in terms of the receivers that they send to; we can find receivers who are similar in terms of who is sending them messages. In each of these analyses, though, we are implicitly treating one mode as "independent" and the other "dependent." The process we are describing, however, is co-evolutionary , with both sending and receiving being dependent. A two-mode analysis would be more appropriate.

To date, there are two main approaches to two (and multi) mode relational data. One approach is to apply a technique of the "correspondence analysis," "singular value decomposition," "multi-modal factoring" type (Faust, 2005). These approaches partition the total pooled variance (e.g. variance across senders in their profile of receivers along with variance across receivers in their profiles of senders). The result is a dimensional decomposition of the variance that can be used to scale both modes simultaneously, and can be used to identify clusters of senders and receivers who are "close" to one another. These are extremely useful outcomes (some examples are given below). Unfortunately, only structural equivalences can be considered – at least in existing software.

The alternative approach is generalized block-modeling (Doreian, et al., 2004). Senders would be classified into partitions based on their profiles of ties to partitions of receivers, and vice-versa. For example, we might identify a partition of message senders who directed communications at all others (spammers), partitions that communicated only with members of their own group, a partition of receivers who did not send, and so on. We might have a prior hypothesis about the number of sending and receiving partitions and the kinds of equivalences that described their relations; or we might explore the data for best-fitting partitions and equivalences. The generalized block-modeling approach provides the greatest fidelity to modeling processes among heterogynous modes of social actors. Unfortunately, existing software is very limited (two modes, small numbers of cases in each mode).

In the next sections, we will provide some examples and some speculations about ways in which casting problems as multi-modal relational networks has been and/or may be of use in understanding data produced by ongoing social processes.

Illustrations of Modality and Equivalence in Social Process Produced Data

Any set of social processes that produce documentation (preferably time stamped!) in the form of transaction records could be treated as a relational data structure, and analyzed using network analytic tools. A good deal of such work has been done, and we are not attempting a survey here. Because of both conceptual and software limitations, we have yet to take full advantage of the approach. A few illustrations will serve to highlight some of potentials and current limitations.

Bibliographic Databases

In his survey article on scientific networks, Howard White (White, forthcoming) demonstrates that the multi-mode, co-evolutionary perspective is becoming the dominant approach in bibliographic studies of academic (mostly scientific, but also some humanistic) communities.

The volume of information that is available in digital form in bibliographic databases is quite stunning and growing very rapidly. One popular resource for literature in biomedicine, popularly known as "Medline" (National Institutes of Health, 2010), currently contains about 19 million citations from a broad range of periodical literature in bio-medical fields. Each record contains authors, titles, abstracts, many full-texts, key-words, venue of publication, date of publication, and other standard fields. A collaborator of the author of this paper has developed software to mine records for additional data (such as the institutional affiliation of authors).

A number of the fields in these data records are very reasonably conceptualized as modes of social actors. Authors and articles are obvious, but important: author-author ties by direct collaboration or citation are stables. When these affiliation networks are examined through time, the rise and fall of article impact, author status, critical paths, and community structure (e.g. how does the size of the giant component evolve?) can be described. Many such analyses exist, though they explore only very small parts of the available data, and rely entirely on structural, rather than regular equivalence notions.

Still to be explored are the effects of other active social agents. Journals and their editors play active roles in shaping the development of fields. Institutions (universities, laboratories, etc.) affect the likelihood of collaboration. Topics (key-words) are combined and re-combined to elaborate existing specialties and stake claims to new leading edges. Emerging empirical work is exploring some of these less traveled paths, and is finding evidence of very complex co-evolutionary dynamics.

Structural equivalence analyses of such multi-modal data would yield particular combinations of authors/venues/key-words/articles that are at particular locations in graphs (high closeness centrality, high betweenness centrality). Regular equivalence analyses would seek to identify parallel and similar structures in, perhaps, varying scientific fields or historical contexts.

Text and Narrative Mining – Integrating Content Analysis with Network Analysis

The method of content analysis is to creating classes of objects (text strings) that have some form of relation with other objects (text strings), and study the pattern for the resulting semantic network. The most obvious and oldest approach is to treat words as objects, and to count the number of times they appear within a defined distance from one another in a text as undirected tie strength. Simple co-occurrence of words is using the notion of structural equivalence. Generally, however, content analysis seeks to create or identify regular equivalence classes. For example, a tie exists if any of the words in the set {pony, horse, pinto,…} are within a given distance of any of the terms {ride, mount,…}.

Commonly, equivalence is imposed by the analyst based on conceptual schema and deep knowledge of the problem. The validity of results, however, depends on the coders and consensus about the dictionary. And, until the dictionary is developed, content analysis of text is slow, somewhat unreliable, and expensive. Processing large volumes of text traffic in anything resembling real time remains a major challenge.

Mining large volumes of texts and multiple coding of the same text to create databases of equivalences is one approach. Google's efforts in developing language translators by building equivalences from multiple translations of the same text, and direct comparisons of web contents (e.g. the same content posted in a web site in German and English) is one feasible approach based on structural equivalence. Alternatively, it might be possible to apply algorithms for identifying approximate regular equivalence classes. Regular equivalence reductions would not yield good textual translations; they would, however, be rather more useful for uncovering meanings and implications of text.

Now consider some complexities. Rather than a single text, suppose that we were working with multiple texts; or considering parts of a text produced by different actors, or texts produced by different actors. Perhaps the texts are "directed" – for example, in a conversation, thread in a discussion board, or email stream. Perhaps, and usually, the texts are temporally ordered.

But imagine if we could define a multi-modal data structure of class(words) by class(words) produced by class(actor) directed to class(actor), at class(time). We can now, potentially, partition the total joint variance, or propose and fit equivalence block models to the entire structure. Why would one? Word prevalence and word adjacency may well be contingent depending on the sending and receiving actors, and may vary systematically as the discourse develops. When texts are examined for examined to attempt to identify unknown authors or their attributes (the writer was raised in the southern United States, for example), multi-modal mining is occurring.

The same kinds of notions of treating parts of texts as objects, and examining them relationally, have been applied to whole narratives. Beginning, perhaps, with the work of Heise (Heise, 1989; Corsaro and Heise, 1990), narratives are treated series of "events" (each of which has affiliated sources, targets, and other attributes), that are ordered by the relations of logically necessary and sufficient conditions for the occurrence of other events. Mining the structure of narratives, identifying logical peculiarities, comparing accounts of the same events by different actors in historical research have generated a (very limited number of) quite interesting results (Griffin, 1993).

Formal analysis of narratives (and the related study of event sequences) have not (to my knowledge) been cast in network-relational terms. Heise's "event" objects, however, can easily be seen as one mode in a relational structure with which authors and targets are affiliated. The structure of narratives as event sequences themselves can be cast as networks, and mined for structural and regular equivalences that would identify characteristic sequences that might vary by author or other affiliated traits.

Cognitive Social Structure

An early, but still very useful, application of multi-mode analysis is that of "cognitive social structure" (Krackhardt, 1987). Data of this type consist of collecting information about the relational structure of a number of objects, as understood by a number of perceivers. For example, the patterns of which persons "liked" which other persons might be reported by each person in a group. The data are three-mode: source of a "liking" relation; target of a "liking" relation; and the rater.

It is possible to examine which raters are similar to which others in terms of the similarities of the "maps" they draw of who likes whom. One could evaluate which actors were similar as sources of liking, based on the profiles of their targets, or (alternatively) based on the degree of similarity in the ratings of this by raters. And so on.

In this example, the sources and targets of liking are two modes of social actors. Even though the two modes contain the same elements, they are not the same mode, because the relation of "liking" is asymmetric. The third mode also has the same index of actors, but is "ratings." My impulse is to treat the "rating" as an "event" – an emergent symbolic or cultural characterization or perception of social structure. This generates a network structure in which k events (where k is the index of group members) each "affiliate" sources and targets of liking. As a structural equivalence problem, we would like to know: which actors are perceived by raters as having similar targets of their liking? Which actors are perceived by raters as being similar in terms of which actors like them? And, which perceivers have similar maps of who likes whom?

"Individual differences scaling," "three-way clustering, " and "multiple correspondence analysis" can be applied to data of this type, if we perceive the questions of interest to be similarity as structural equivalence (e.g. Arabie, et al. 1987). One might seek a further reduction of the modes into regular equivalence categories: are there "kinds" of sources of liking relations who have different spectra across "kinds" of targets of liking, as perceived by "kinds" of perceivers.

Virtual Communities: Email, Blogs, WWW, Social-Networking, Netgames, Open-source Development Communities

Social processes occurring in "new media" leave logs of transactions. Because transaction records are already in forms that are fairly easily machine processed, and because the volumes of data available are huge (by orders of magnitude compared to old-media documents) a great deal of effort is currently directed toward their analysis. The largest part of the effort to exploit data sources of this type so far has been by information scientists, researchers in complex network dynamics, and similar fields. A large part of this work has treated the data as networks, and has applied network analysis tools (often from engineering and physics, more than social sciences). Much of the effort has focused on problems of search, robustness, and other aspects of network topology. Some work has been done on more traditionally sociological topics such as identifying communities, core-periphery structures, central nodes, and the like. The two key concepts that we've explored at some length in this paper – modality and equivalence – have not yet been extensively applied in new media studies. Here are a few ideas of how they might be.

Email, phone, discussion board, texting, and blog transactions are routinely archived in digital form. These "traces" reveal the structure (and often the content, as well) of very large volumes of one-to-one and one-to-many communications among social actors. Typically the records exist in transactional form with fields that record information that can be treated as multiple modes, and attributes.

An email object, for example, has a source, one or more destinations (of various types), a subject line, often some indication that it is part of a thread (RE, FWD), and a text. The time and location from which it was sent, the path it followed, and the content and attachments are often available. Senders and receivers are "affiliated" with messages (and form either regular or structural equivalence classes). The text mining of subject lines and/or message texts can produce regular or structural partitions that are symbolic/cultural contexts within which senders, receivers, and particular sets of messages are affiliated (again, either regularly, structurally, or both). The time-stamping of traffic could, in principle, allow the characterization and analysis of the "shape" of the multi-modal space, and the characterizing the trajectories (direction and speed) of topics, sources, and senders.

Social networking sites, url-url linking in the WWW, logs of games, and open-source software development communities are some examples of virtual communities that are self-selecting affiliation structures, logically parallel to "voluntary associations" in traditional social science studies (Cress, et al. 1997). Virtual communities are many-to-many structures that are created by affiliation, and have a bi-partite network structure. They may also embed direct connections between individual agents and direct connections and or affiliations among event/symbolic/organizational social agents. Some work is developing in this space, using network approaches. The notions of modes and equivalences may provide some interesting new directions. Two very brief speculations:

Open-source software development processes and communities have been studied (primarily by computer scientists), in part because of the large quantity of high quality data produced by the documentation of such collaborations (e.g. Sourceforge, 2010). Participants affiliate with one or more projects, taking roles in creating, revising, and assembling components of software programs. Within a project, actors affiliate with one or more components, which are themselves "affiliated" with other components (code segments call other code segments). The entire structures evolve over time, driven by the internal logic of the task, but also by the social logics of leadership, status seeking, cooperation and altruism. We should note that the vast majority of open-source software projects are failures – both technically and socially. Structural analysis enables us to understand particular projects; regular analysis could provide more general insight into the commonalities of successful and failed communities.

Multiple user interactive games such as Warcraft (Warcraft, 2010) are hosted on servers that log all transactions. Some of the communities are huge (millions of participants), and the transaction logs are almost incomprehensibly large. These communities (like social networking sites) are of interest both in themselves as new social phenomena, and because they are naturally occurring experiments in network dynamics, exchange dynamics, and other structured interactions. Games are particularly intriguing because human participants may construct one (or more) identities, create and affiliate/disaffiliate with both long term (e.g. "kingdom") and short term (e.g. "quest") symbols and organizations. The symbolic and organizational classes evolve by both affiliation and by selection dynamics within their own mode (two "quests" may join forces to fight a battle). The structural equivalences of these multi-mode structures may be important for the information they provide us about evolving network topologies. Regular equivalence structures might tell us something deeper about more general patterns and dynamics by which communities and more complex social structures are constructed and de-constructed.

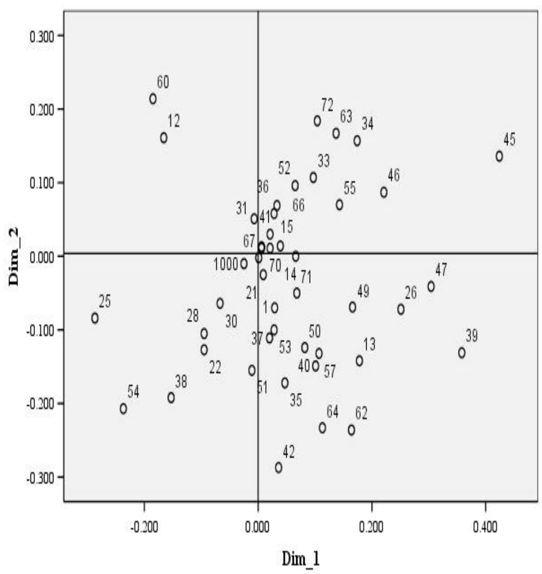
Policy Networks and Politics

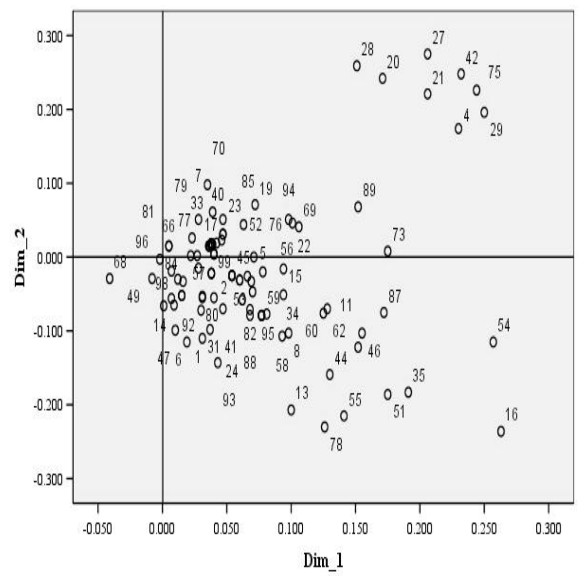
The relational network perspective has, particularly recently, been advancing rapidly in political science and political sociology. It is quite easy to see records of political acts (e.g. voting, making donations) from a relational network perspective. The votes of citizens for candidates and initiatives accumulate over election cycles; the votes of legislators on bills accumulate over a session; the votes of justices on courts accumulate over time.

The more traditional approaches to such archival data are to use attributes of actors (donors, voters, legislators, judges) to predict their orientation toward (or affiliation with) particular outcomes (candidates, bills, court cases). Increasingly, however, political analysts have become more sensitive to the non-independence of these events across actors and time. The relational perspective can provide some different insights to such complex processes than conventional statistical approaches.

Bowler and Hanneman (2006) examined the data archive collected by the Secretary of State of California on donor's contributions for and against 59 ballot initiatives over the period 2000-2004. Donors and initiatives can be cast as two modes of social actors that are brought together into a co-evolving relationship by the act of donations. Donors who support the same sets of initiatives come to perceive themselves as a "community" or "social movement." The initiatives that are supported by coherent sets of constituents are perceived to be part of larger policy issues or ideologies. Past collaboration may breed future cooperation among donors; as multiple initiatives become seen as part of an agenda (e.g. California's "Proposition 13" and the "tax revolt"), they may spawn new initiatives. Figure 3 shows a mapping of major donors (those who gave more than $1M US to more than one initiative), and a mapping of the initiatives in the joint "policy space."

Figure 3. California Ballot Initiatives (left) and Major Multi-Campaign Donors (right) in Joint Space.





The analysis suggests both dimensionality (the authors interpret the dimensions as liberal/conservative and statist/anti-statist), and clustering (e.g. labor unions and Democratic political groups often co-donate). This analysis pools across several election cycles. Treating each election as an additional "mode" – a contextual event, might improve the analysis, and would be able to show how the policy-space and the donor-space co-evolved. The analysis also relies on structural equivalence (donors are similar to the extent that they had same profile of ties across 59 initiatives; the initiatives are similar to the degree that they were supported by exactly the same individual donors). Greater insight might be possible by seeking patterns of regular equivalence – "types" of donors based on similarities in their profiles of support for "types" of initiatives, and vice versa.

Organizational Ecologies

I have a particular interest in organizational ecology – an attempt to apply principles and theories of human ecology (Hawley, 1986) to populations of (usually formal, but sometimes voluntary, organizations. Glenn Carroll and Michael Hannan (Carroll and Hannan, 2000) are perhaps the key figures with regard to formal organizational ecologies, McPherson (Cress, et al. 1997) is the leading figure with regard to organizational ecologies of voluntary associations.

The core idea here is that organizations that perform particular specialized functions locate in non-random ways in human settlements. One reason that they make non-random choices is the presence of other functions in particular places. Places also have independent attributes that make them differentially attractive for different organizational functions (e.g. they are located on a river).

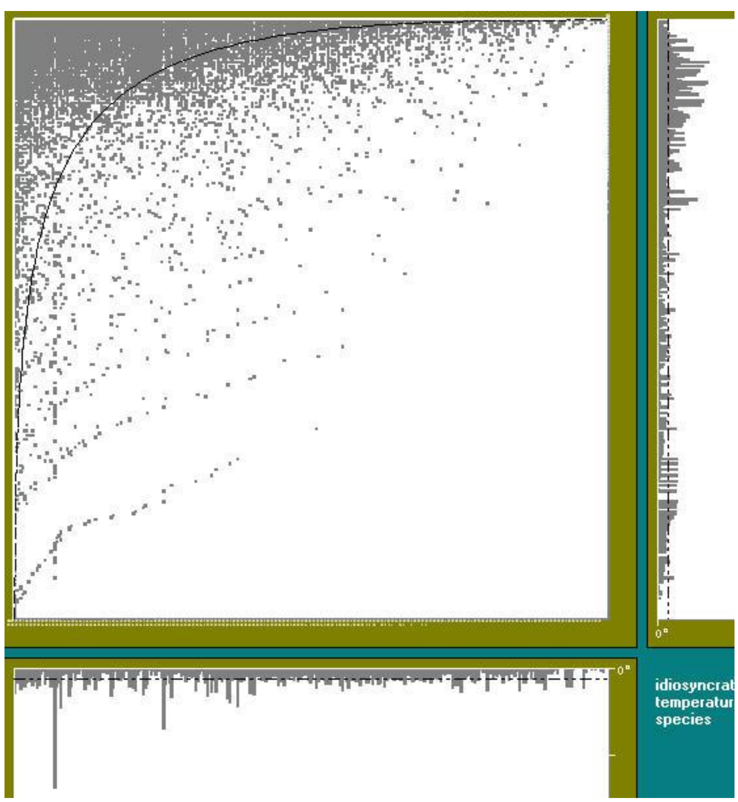
One supposes that one may identify patterns of organizational density that define "types" of organizational communities. One may also identify "types" of places that select for varying mixes of organizational types. That is, populations of organizations and populations of settlements "co-evolve" by the processes of affiliation (birth, death, change in function, migration). The analogy to biological ecology is extremely strong, so the notions of modality and equivalence could easily be applied to biological and ecological co-evolutionary processes.

Business directories and listings are routinely produced as adjunct to processes of marketing (e.g. finding all the dentists in Omaha, if you have a new dental instrument to sell). Directories such as Reference USA (2008) list about 13 million enterprises, and give a number of attributes of them (primary product, approximate sales volume, location, etc.). The data are stored in a relational form, with some classes or primary keys: location, primary product).

Individual establishments are affiliated with particular communities, and also with the social construct of a "primary product" or "industry." Individual establishments have attributes (size, ownership form, etc.) that may be used as partitions or colors. Communities, as well, have attributes (e.g. connection to logistics networks, population size, political centrality) that may partition them and shape affiliation processes. Industries have theoretically important attributes (e.g. capital cost barriers to entry, scope of market, location in commodity chains) that may color their dynamics.

One theory of organizational and community ecology hypothesizes a "central place" hierarchy, in which (possibly following a power-law distribution) both functions and places form nested hierarchies. Figure 4 shows a display of this joint hierarchy for the state of New Mexico in 2004.

Figure 4. Organizational/Community Nestedness in New Mexico, 2004.



Source: Data from Reference USA. Analysis by the author using NestCalc (Atmar and Patterson, 1993).

It is clear from the nestedness diagram that a simple scale-free network does not apply to the joint affiliation. A structural block model might do at least as well in fitting the data. More interesting, however, would be a regular reduction of the data: are their substitutable sets of organizations that are present in varying combinations across multiple, qualitatively different types of communities?

The World Systems of Nation-States: Multi-plex Relations

International relations research (primarily in the disciplines of political science, strategic studies) and "world systems" research (primarily in sociology, economics, and political science) seek to characterize and understand changing patterns of relations among nation-state entities. There are huge literatures and great diversity in theoretical perspective, analytic goals, data sources, and methods. For our current purposes, my "take" on the field is very similar to that taken in a widely cited and controversial article by Snyder and Kick (1979).

Snyder and Kick work with a data structure that can be thought of as a relational database composed of nation-states that have a variety of orientations toward and relations with other nation-states. The "orientations toward and relations with" are composed some asymmetric relations (e.g. origin nations engage in military interventions to destination nations, origin nations export commodities to destination nations, and origin nations send diplomatic representatives to destination nations. Nations are also tied symmetrically by being co-signatories to treaties. The first three of these relations are direct ties between two modes (origin and destination nation) that have the same class members. The "treaties" tie is actually an "affiliation" – where nations are "nested" within supra-national "agents" (treaties).

Let's keep it somewhat (but not too!) simple, for current purposes. The data array could be seen as consisting of three asymmetric square two-mode tables (origin by destination nations) containing information on diplomatic ties, trade, and military interventions. There is also a fourth data table here, which is an affiliation matrix of nations by treaties. Snyder and Kick took the approach of inducing a nation-by-nation co-participation in treaty relation from the affiliation data, producing four matrices of the same dimension.

Writing 30 years ago, Snyder and Kick greatly advanced the sophistication of world system analysis by including multiple kinds of ties between their origin and destination nations simultaneously. They innovated by executing a block model of structural equivalence of the nations that included all of the variance in trade, military interventions, diplomatic exchange, and treaty co-participation. They used the often criticized (but remarkably robust!) indirect-method CONCOR algorithm to classify nations into 10 approximately equivalent blocks. That is, they performed a blocking based on structural equivalence across four relations, simultaneously. For the time, it was a remarkable achievement. The results have proven to be rather robust. The sophistication of the analysis has rarely been matched.

Still, my current argument suggests a somewhat different approach to the problem. Snyder and Kick chose to take the treaty co-participation relation and use it to induce a nation-nation relation, rather than conceptualizing treaties as emergent "social facts." The type of equivalence chosen in their analysis was structural – that is, two nations would appear in the same block of nations with a probability proportional to the similarity of their profiles of ties all other specific nations (across the four types of ties). A somewhat more abstract, and possibly more theoretically useful, approach would be to treat the problem as one of generalized regular equivalence: two nations fall in the same class to the extent that they are related in the same way to at least one member of each other class. Puerto Rico is dependent on the United States, and Algeria on France. The U.S. and France are not a structurally equivalent class, but they are a regularly equivalent class. Furthermore, there is no logical necessity that the classification of nations as senders of a particular type of tie produces the same blocking as the classification of nations as a recipient of a particular type of tie.

If we did pursue this problem as a general structural equivalence problem, we might proceed by performing a multiple correspondence analysis, co-factoring, or co-clustering to produce mappings of the joint variance in all the relations. Depending on the scaling, this approach would provide us a decomposition of the total variance among origins and destinations, trade, diplomatic ties, military interventions, and treaties. We could assess the dimensionality of the joint variance, and locate origin nations, destination nations, treaties, trade, intervention, and diplomatic relations in joint space. This would allow us to see, for example, which pairs of nations were closest in "treaty space", and whether these clusters of nation-treaties were also close to export flows, import flows, and diplomatic relations.

Approaching the problem as a generalized block model, one would proceed rather as Snyder and Kick did, but one could impose regular equivalence on some relations (say trade), but structural equivalence on other relations (say treaty ties).

Trade Dynamics in World Systems

Patterns in volumes of trade in commodities among national economies are of interest for a number of national strategic, economic and trade policy, and social science theoretical reasons.

The data are stored in a relational database compiled by International Monetary Fund , compiled from national government's reports, and surveys. The data describe (aggregated by year) the volume of flow from each of a large number of nations, to each of a large number of nations, of each of a large number of commodities. The basic relational structure is a four-dimensional many-to-many relation: each sending nation may send volumes of many commodities to many receiving nations at many points in time. The modes here are exporters and importers, and they are connected by a crossed relation of commodities and time.

From a blocking or clustering perspective, we are interested in identifying (or modeling) sets of exporting nations that are regularly equivalent with regard to importing nations; importing nations that are regularly equivalent with regard to exporting nations. We also want to know what commodities are regularly equivalent. We might hypothesize that some nations at some points in time are producers and consumers of high tech goods, for example. And, we are interested in seeing whether nations change roles as importers and exporters of various types of commodities over time. The analysis then involves the trajectories of equivalence classes of importers, exporters, and commodities in time.

An exemplary analysis of trade tables very similar to this description was accomplished by Smith and White (1992). Proceeding from a world-systems perspective, they sought to identify blocks of nations (at each of three points in time, separately), that were regularly equivalent in exporting 15 commodities (chosen as indicators of core commodities from a prior factor analysis of a large number of commodities that identified five dimensions of commodity flows). Smith and White proceeded by producing measures of regular equivalence for pairs of nations across the 15 commodity flow tables simultaneously, and then used block modeling to identify five blocks. The dynamic dimension was studied by looking at mobility of individual nations from one regular block to another between time 1 and 2, and between time 2 and 3.

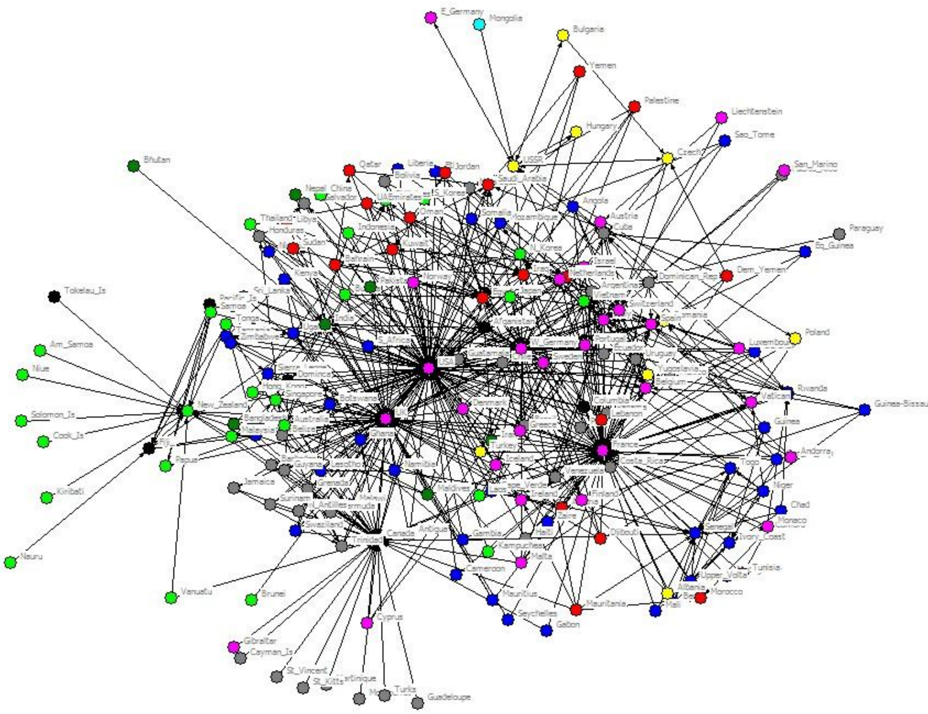
The approach of Smith and White identifies two modes (exporters and importers) as social actors, and sees them as having 15 relations at each of three points in time. We could just as easily treat this as 45 relations. A fully simultaneous blocking of the data would allow that blocks of exporters might have different members than blocks of importers, that the 15 commodities could be blocked into a smaller set of classes, and that these blockings might change over the three time periods.

The notion of treating exporters (or originators of transactions) and importers (receivers of transactions) as separate modes – allowing that the variance of each mode might contribute different amounts to the total variance – has recently been pursued by Boyd et al. (2010). These authors, rather than seeking regularly equivalent blocking (as did Smith and White), fit a core-periphery model (a blocking with high density of ties among members of the core, low tie density among members of the periphery, and agnostic about ties in the off-diagonal blocks), using structural equivalence.

Patterns of International Student Mobility for Higher Education

In an undergraduate honors seminar, Hiroko Inoue and I were examining the changing structure of flows of higher education students abroad to study. Data are collected on this subject by the United Nations (cite), and presented as simple tables showing the number of students flowing from nations to other nations. The volumes of the flows are aggregated annually, and are available in the database for only the five largest outflows from each nation. Figure 5 shows one time-slice (1976). In this figure, links are shown if more than 5% of the students from a nation go to the destination. The nodes are colored by geographical region, but arrayed in two-dimensions of the similarities of their relational profiles.

Figure 5. Destination Flows of International Students, 1976.



We had a number of questions of interest in this analysis; in addition to whether the patterns had changed between the two time-points (we suspected that the patterns had, with the collapse of the Soviet empire, the rapid expansion of total volume, and the development of high quality universities outside of the European-North American axis. We believed, among other things, that nations that were former colonies were likely to have large flows to former metropolis; that nations that were in the same "civilization region" were likely to have high flows among them; that nations which were in the core/semi-periphery/and periphery in trade-flow terms were likely to display the same kind of hierarchy in flows of students.

A fairly straight-forward approach to the problem would be to treat the data as a three-dimensional array (origin, destination, and year). There would be two modes of social actors (origin nation, destination nation) and two relations (flows in 1976, flows in 2005). Each of the objects in each mode could be characterized by attributes such as "former British colony" "Slavic nation" "core nation." One could then partition or block the data on one or more of these attributes, and examine block densities.

It is logically possible to treat these attributes as relations of affiliation. Two nations that fall within the same "civilization" region may be thought to be "affiliated" with the sociologically meaningful "identity." "Civilization" regions may be thought to have emergent relations *sui generis* that may be time-variant (as in current tensions between Christian and Muslim "worlds").

Thinking about some of the "attributes" of nations, in this case, suggests that they might be usefully re-cast as modes, rather than as "independent" replications of a level of treatment of predictor or independent variable. We are then led to ask some questions that we would not have otherwise; for example, how similar or "close" former British and former French nations to each other in terms of their patterns of flows at both points in time?

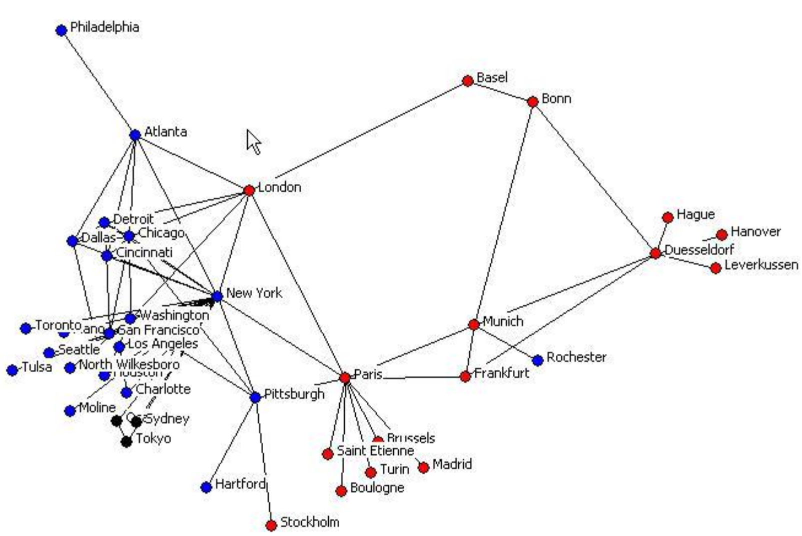
Global Command Centers of Capitalism: Headquarters Cities, Board Overlaps, Industrial Ecologies.

Another main line of work in globalization and world-system studies focuses on the relations among social classes, multi-national organizations, production chains, and spatial location. Very large corporate enterprises (e.g. the Fortune global 500), which are dominant players in high-value product industries (through networks of subsidiaries and value production chains) are headquartered in particular places (cities, nations, regions), and are synchronized by the information flowing through interlocking corporate boards. While multi-national corporations are often seen as the main locus of agency, organizational fields, corporate directors, urban spaces (Sassen, 2001), and geographical/spatial diversity are all seen as playing some role in shaping the effects of one another (Neal, 2008).

Among the main empirical questions in this line of work is the relative importance of each of the modes in shaping the overall distribution of global economic activity, the location of "leading edges" and "central actors" in the system, and the trajectories of cities, firms, industries, and directors.

There are a number of modes in this problem (firms, industries, cities, directors), as well as a variety of relations that define the overall variance. Firm headquarters are affiliated by the cities in which they are contained; directors are connected by virtue of the firms on whose boards they serve; cities may be similar to one another by virtue of the industries of the firms that are headquartered in them, and so on. One map of part of these relations is shown as figure 6.

Figure 6. Headquarters of Fortune Global 500 Corporations and Board Interlocks



Source: Kentor and Hanneman, 2008.

In this figure, cities are identified as connected if they are containers of global 500 corporations that have overlapping boards of directors. The spatial layout of the display is defined by the first two dimensions of the city-city distance matrix induced from the city/corporate board interlock relation. The world regional locations of cities are shown as coloring. New York and London appear as hubs of the first order (Tokyo, contradicting some observer's hypotheses, does not), and Paris as a hub of the second order, nested largely within Europe.

This illustration is very basic and does not include all of the modes directly. The addition of even a second time point (work in progress) would enable the calculation and analysis of the trajectories of objects and their clustering. And, the approach considers only structural distance or equivalence; much world system theory (for example, the work of Sassen, 2001) actually proposes regular equivalences in such notions as "global cities" as hubs that may organize somewhat different parts of global production, but do so in equivalent ways.

Structure from Lists: A Faculty-Student Network

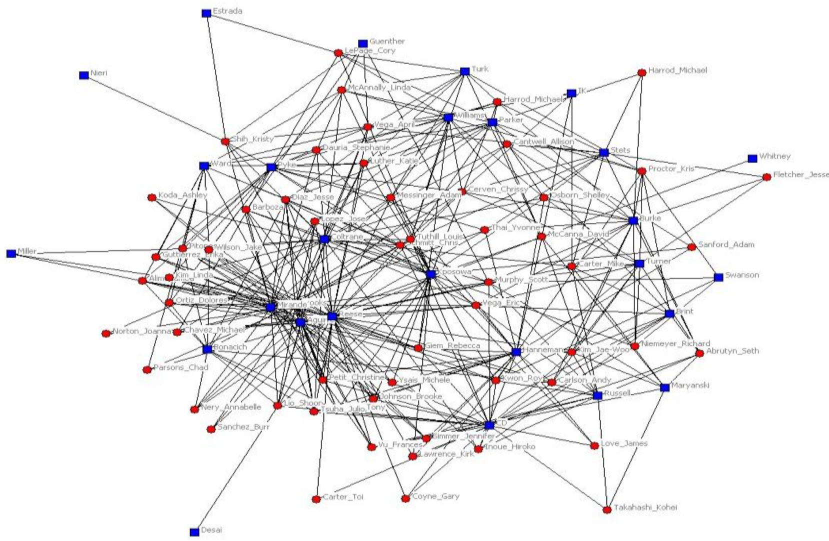
Many archives generated by social processes are assembled as accumulating lists of transactions. In the graduate program at my university, every student prepares an original research paper under the guidance of (usually) two to four faculty, as a requirement in the master's degree program. The archive consists of a list of students, the paper title, date completed, grade received, and the names of the faculty members of the reviewing committee.

It is very easy to re-cast this list as a relational data structure. One approach is view the structure as composed of relations among three modes: faculty, students, and papers. The papers have the attributes of the date completed and grade received – these are not types of social actors (of course, grades and dates could be treated as modes). The faculty and the students are classes of individual persons; the papers are "events" with which students and faculty are affiliated. Each student can be affiliated with one and only one paper; many faculty members can be affiliated with many papers. As a practical matter, because of the identity relation between students and papers, this becomes a two-mode problem.

Here we have two classes of individual persons, and one class of events. The sociologist's natural inclination is to see this as a macro-micro problem: how do individual persons (faculty and students) interact to create "emergent" structures (papers)? And how do "events" create ties among actors? In this case, the question then is: how similar are students based on their profile of ties with faculty (through the context of the research paper)? How similar are the faculty based on their profile of ties with students (through the context of the research paper)?

Figure 7 shows a map, in two-dimensional space, of the two-mode faculty/student network. The graph locates the nodes in the first two (non-trivial) dimensions of a singular value decomposition of the two-mode matrix (students by faculty, with cell entries representing whether a faculty member was on the evaluation committee of a student.

Figure 7. Two-Mode Structural Equivalence SVD Mapping of Faculty-Student Affiliation in the UC Riverside Master's in Sociology Program.



The similarities in this example are based on structural equivalence. Viewing the map, there is some apparent clustering of faculty and of students – a dense "core" of faculty and associated students (lower left) and a more dispersed and varied "periphery." One might push the analysis further by relaxing the equivalence to a generalized block structure. A model might be proposed consisting of "core" and "periphery" partitions of both faculty and students. In the model, the block structure would be: core faculty are tied to all other core faculty; core faculty are tied to at least one core student; core faculty may are tied to one or more peripheral faculty and to one or more peripheral students, etc.

A two-mode variance decomposition allows us to examine what portion of the total variance is due to dissimilarity among faculty profiles, and dissimilarity among student profiles (are faculty positions clear and orderly, but students less coherent?). Variance decomposition also allows us to assess the dimensionality of the pooled variance (is a map in two-space, as shown above, a reasonable representation?). To date, this approach is limited to structurally equivalent relations, however. To pursue the analysis of different "social roles" among faculty and among students (which require measures of similarity based on regular equivalence, or a mix of equivalence types), block-models would be applied.

Conclusion

We have argued that many social processes produce records of transactions, and this is particularly true of digitally-mediated interactions. Conceptualizing the social processes that produce these records as co-evolutionary processes of relation-making among and within heterogeneous agent classes (modes, in network analysis terms) may be helpful in structuring the flows of information into useful data. Analytic tools for working with relational data may be powerful approaches to mining such data structures. In particular, the strategy of identifying similarities in relational patterns within and between agent classes on the basis of structural and regular equivalence can produce useful insights into complex and voluminous data.

Many “structural” social scientists have ways of thinking about social-process produced data that are highly compatible (at a broad level) with the ways that computer and other information scientists think. The notions of “objects” containing data structures and “methods” are highly compatible with the notions of “social actors” and “ties.” Both structural social scientists and information scientists tend to view phenomena as complex, co-evolving, relational processes. There is a great deal in common to build on as we approach the study of the increasingly large volumes of process produced digital data that document more and more (but not all!) of social life.

Two conceptual distinctions that are of great importance in social network analysis – the notions of modes and types of equivalences – may provide bridges between the skills and expertise of computer and information scientists with data structures, and the knowledge that social scientists have of the processes that produce the data. A number of examples have been briefly explored of data produced by social transactions, and how they might be (or in some cases have been) approached using modes and equivalences. The work that has been done is primitive when compared to the potentials. There is a great deal more that could be done in the domains that we’ve briefly explored, and in many others. To exploit this potential, increasingly close collaboration between social and information sciences will be necessary.

References

Arabie, Phipps, J. Douglas Carroll, and Wayne S. DeSarbo. 1987. *Three-way Scaling and Clustering*. Sage Publications.

Atmar, W. and B. Patterson. 1993. "The Measure of Order and Disorder in the Distribution of Species in Fragmented Habitat." *Oecologia*. 96:373-382.

Batagelj, Vladimir, Andrej Mrvar, Anuška Ferlioj, and Patrick Doreian. 2004. "Generalized Blockmodeling with Pajek" *Metodološki Zvezki*, Vol. 1, No. 2, 455-467.

Bowler, Shaun and Robert Hanneman. 2006. "Just How Pluralist is Direct Democracy? The Structure of Interest Group Participation in Ballot Proposition Elections." *Political Research Quarterly*. 59(4): 557-568.

Boyd, John P., William J. Fitzgerald, Matthew Mahutga, and David A. Smith. 2010. "Computing continuous core/periphery structures for social relations data with MINRES/SVD." *Social Networks*, 32, 125-137.

Breiger, R.L. 1974. "The Duality of Persons and Groups." *Social Forces* 53:181-190.

Breiger, R.L., S.A. Boorman, and P. Arabie. 1975. "An Algorithm for Blocking Relational Data, with Applications to Social Network Analysis." *Journal of Mathematical Psychology*, 12.

Carroll, Glenn R. Michael T. Hannan. 2000. *The Demography of Corporations and Industries*. Princeton, Princeton University Press.

Chen, Chomei. 2006. "CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature." *Journal of the American Society for Information Science and Technology*. 57(3): 359-377.

Chung, Fan R.K. 1997. *Spectral Graph Theory*. American Mathematical Society.

Collins, Randall. 1998. *The Sociology of Philosophies: A Global Theory Intellectual Change*. Cambridge, MA: Harvard University Press.

Corsaro, William and David Heise. 1990. "Event Structure Models from Ethnographic Data." Pp. 1-57 in Clifford Clogg (Ed.), *Sociological Methodology 1990*. Cambridge MA: Basil Blackwell.

Cress, Daniel M., J. Miller McPherson, and Thomas Rotolo. 1997. "Competition and Commitment in Voluntary Membership: The Paradox of Persistence and Participation." *Sociological Perspectives*. 40(1):61-79.

Doreian, Patrick, Vladimir Batagelj, and Anuska Ferligoj. 1994. "Partitioning Networks Based on Generalized Concepts of Equivalence." *Journal of Mathematical Sociology* 19(1):1-27.

Doreian, Patrick, Vladimir Batagelj, and Anuska Ferligoj. 2004. "Generalized Blockmodeling of Two-mode Network Data. *Social Networks*. 26:29-53.

Everett, Martin G. 1994. "Regular Equivalence: General Theory." *Journal of Mathematical Sociology* 19(1):29-52.

Faust, Catherine. 2005. "Using Correspondence Analysis for Joint Displays of Affiliation Networks." Pp. 77-97 in Peter J. Carrington, John Scott, and Stanley Wasserman (Eds.). *Models and Methods in Social Network Analysis*. Cambridge: Cambridge University Press.

Griffin, Larry J. 1993. "Narrative, Event-Structure Analysis, and Causal Interpretations in Historical Sociology." *American Journal of Sociology*. 98:1094-1133.

Hanneman, Robert A. and Mark Riddle. 2005. *Introduction to Social Network Methods*. Riverside, CA: University of California, Riverside (published in digital form at http://faculty.ucr.edu/~hanneman/)

Hawley, Amos. 1986. *Human Ecology: A Theoretical Essay*. Chicago: University of Chicago Press.

Heise, David. 1989. "Modeling Event Structures." *Journal of Mathematical Sociology*. 14:139-169.

Institute for Scientific Information (ISI). 2010. *ISI Web of Science*. Thompson Reuters.

International Monetary Fund. (annual). *Direction of Trade Statistics*. Publications Services, I.M.F. Washington, D.C.

Kentor, Jeffrey and Robert Hanneman. 2008. "The Integration, Network Roles, and Economic Functions of Global Cities." Paper presented at the *annual Sunbelt Meetings of the International Network of Social Network Analysts*. Orlando,FL.

Krackhardt, David. 1987. "Cognitive Social Structures." *Social Networks* 9:104-134.

Lorrain, F. and White, H.C. 1971. "Structural Equivalence of Individuals in Social Networks." *Journal of Mathematical Sociology*, 1, 49-80.

Neal, Zachary P. 2008. "The Duality of World Cities and Firms: Comparing Networks, Hierarchies, and Inequalities in the Global Economy." *Global Networks* 8(1):94-115.

Newman, M.E.J. 2006. "Modularity and Community Structure in Networks." *Proceedings of the National Academy of Sciences of the United States of America*. 103(23):8577-8582.

ProQuest. 2010. *CSA Sociological Abstracts*. Ann Arbor, MI.

Reference U.S.A. 2008. *Reference U.S.A. U.S. Businesses Database* (http://www.referenceusa.com/)

Sassen, S. 2001. *The Global City* (2nd ed.). Princeton, N.J.: Princeton University Press.

Scott, John. 1991. *Social Network Analysis: A Handbook*. London: Sage.

Smith, David A. and Douglas R. White. 1992. "Structure and Dynamics of the Global Economy: Network Analysis of International Trade 1965-1980." *Social Forces* 70(4): 857-893.

Snyder, David and Edward L. Kick. 1979. "Structural Position in the World System and Economic Growth, 1955-1970: A Multiple-Network Analysis of Transnational Interactions" *American Journal of Sociology* 84(5): 1096-1126.

Social Network Analysis and Mining. 2010. *Journal Website*. http://www.springer.com/springerwiennewyork/computer+science/journal/13278

Sourceforge. 2010. *Find and Develop Open-Source Software*. http://Sourceforge.net/

United States. National Institutes of Health and United States National Library of Medicine. 2010. *PubMed Home*. http://www.ncbi.nlm.nih.gov/pubmed.

Warcraft. 2010. *World of Warcraft Community Site*. http://www.worldofwarcraft.com/index.xlm

Wasserman, Stanley and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Watts, Duncan J. 2003. *Six Degrees: Science of a Connected Age*. NY: W.W. Norton.

Webb, Eugene J., Donald T. Campbell, Richard D. Schwartz, and Lee Sechrest. 1966. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally.

White, D.R. and Reitz, K.P. 1983. "Graph and Semigroup Homomorphisms on Networks of Relations." *Social Networks*, 5, 193-234.

White, Howard. Forthcoming. "Scientific and Scholarly Networks." In Peter Carrington and John Scott (Eds.). *Handbook of Social Network Analysis*. London: Sage.

Wikipedia. 2010. (http://en.wikipedia.org/wiki/Object_oriented_language (May 3, 2010).