Joslyn, Cliff A: (2002) "Network Worlds: From Link Analysis to Virtual Places", in: Proc. 2002 Conf. on Virtual Worlds and Simulation, SCS Western Multiconference, Society for Computer Simulation, San Antonio

Network Worlds: From Link Analysis to Virtual Places*

Cliff Joslyn †

March, 2002

Abstract

Significant progress is being made in knowledge systems through recent advances in the science of very large networks. Attention is now turning in many quarters to the potential impact on counter-terrorism methods. After reviewing some of these advances, we will discuss the difference between such "network analytic" approaches, which focus on large, homogeneous graph strucures, and what we are calling "link analytic" approaches, which focus on somewhat smaller graphs with heterogeneous link types. We use this venue to begin the process of rigorously defining link analysis methods, especially the concept of chaining of views of multidimensional databases. We conclude with some speculation on potential connections to virtual world architectures.

1 Network Analysis of Knowledge Systems

There has been a great deal of recent work within the knowledge systems community concerning the role that graph and network theory can play in the organization and retrieval of information. The basic idea is that knowledge systems, including document corpora, but most especially the World Wide Web, can be amply represented as a network or graph-theoretical structure. Then, mathematical and statistical properties of these graphs can aid in locating information. While these developments stem from random graph theory [4], they try to identify non-random structures in naturally occuring network structures.

Perhaps the most famous of these recent mathematical discoveries are so-called "small world" properties [18, 23]. Derived from the famous "six degrees of separation" phenomena from social network theory, they hold when a locally clustered structure is complemented by "short cuts" which produce surprisingly short average path lengths through the graph. Small world properties have been observed in many networks, including biological, socio-technical, document corpora, and the Web [24].

Other statistical properties of large graphs are also invoked, including power law distributions [1] and Kleinberg's "authoratative sources" method [9]. In turn, there has been research on how to exploit these properties for retrieval and navigation [15] and other applications.

Together we will call such approaches "network analytical". There's currently much excitement over them, to the point where it's fair to say there's a bit of a "bandwagon" effect. We almost want to ask what's not a graph? What doesn't follow a power law?

*Presented at the 2002 Conference on Virtual Worlds and Simulation, San Antonio.

†Distributed Knowledge Systems and Modeling Team, Modeling, Algorithms, and Informatics Group (CCS-3), Mail Stop B265, Los Alamos National Laboratory, Los Alamos, NM 87545, USA, joslyn@lanl.gov, http://www.c3.lanl.gov/~joslyn, (505) 667-9096.

For example, there is interest in bringing these ideas to bear on counter-terrorism applications. Shortly after the September 11 terrorist attackes, Stewart compiled a network analytical representation of the terrorist cells [22]. The left side of Fig. 1 shows the results, with the strength of connection indicated by the thickness of the link.

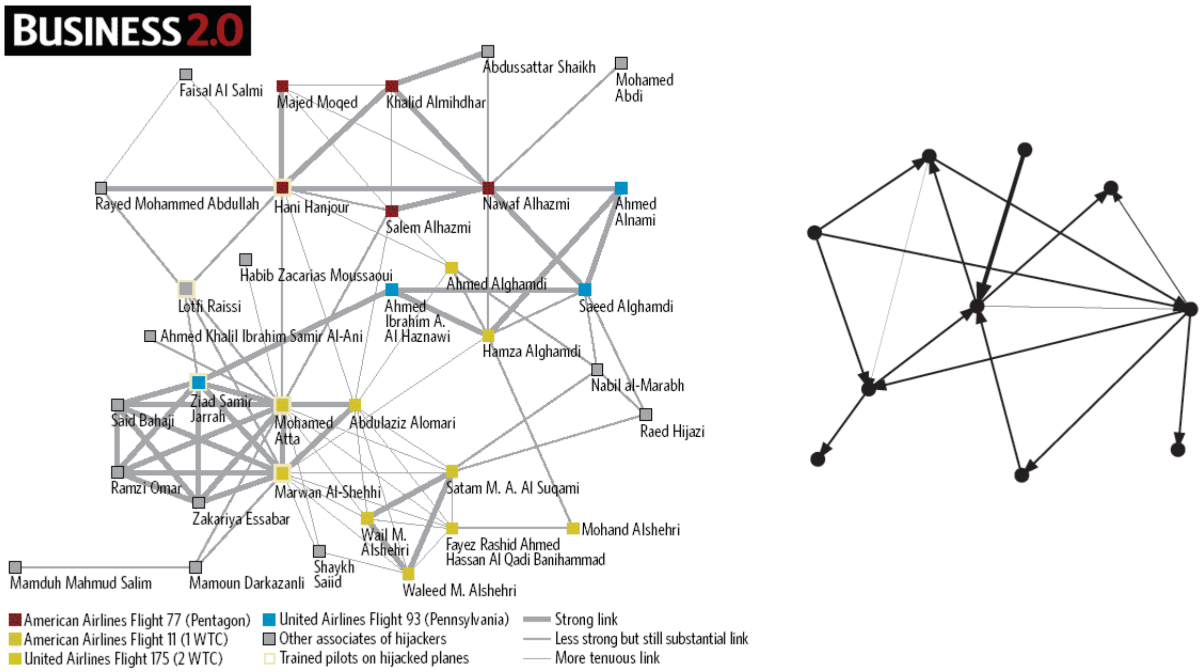

Figure 1: (Left) "The Small World of Mohamad Atta" [22]. (Right) Typical network analysis data structure.

What characterizes these network analysis methods is:

- A single directed or undirected graph.
- The possible use of weightings to indicate the strength of connection.
- Methods applicable to very large graphs, on the order of 10^6 to 10^9 nodes.

Data sources for such structures include large single relations or tables, typically represented as matrices. Such structures are also closely isomorphic to standard modern Web hypertext protocols such as HTTP. The right side of Fig. 1 shows a possible representative small case.

2 Canonical Representations and Typed Link Networks

In prior work [12], we have sought to identify canonical, mathematically sound data structures which are minimally sufficiently complex to serve multiple purposes in knowledge systems, including representation of knowledge networks, dcoument and hypertext corpora, ontologies and other semantic structures, user communities, and virtual world architectures. We have proposed **semantic hyperwebs** as hierarchical lattices of **weighted, directed, labeled, hypergraphs** as being such structures [12], and asserted that they are quite similar to Sowa's conceptual graphs [21].

One way in which semantic hyperwebs differ from the network analytical structures mentioned above is that as hypergraph based they allow representation of relations among more than two things: general n-ary relations as n-ary links, or graph edges.

However, for our purposes here we wish to discuss their nature as *labeled* graph structures. Mathematically, we are invoking the structures alternately called multi-graphs, labeled graphs, or graphs with typed or colored links. Where mathematically networks map to a simple binary relations, labeled networks map to a union of such relations.

As shown in Fig. 2, such structures have **heterogeneous link types:** items can be connected in more than one way, and any two items can be connected by more than one link. So unlike simple networks, such labeled systems can represent structures which Carley [5] identifies as **meta-networks**, or networks of networks. These are useful in representing socio-technical networks in particular, as shown by Barret *et al.* [3], who use link types to indicate transportation modalities.

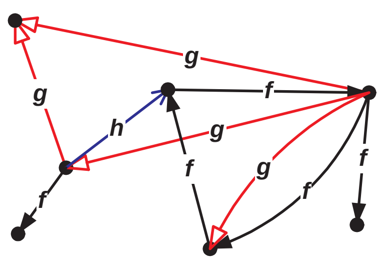

Figure 2: An example of a typed-link network.

It's significant to note that such labeling is the minimal requirement for representing **semantic information** in systems, for example in semantic network and ontological structures [11]. Link types then correspond to types of semantic relations, indicating the qualitative relationships among nodes, which are central both to computer-based knowledge representation and natural language approaches [6]. In the context of knowledge networks in particular, such structures correspond to typed-link hypertext networks. Such models were advanced in early hypertext standards [17] and elsewhere [2].

Consider a typical database of interest in counter-terrorism, for example listing known individuals and some of their characteristics, perhaps including name, address, aliases, age, birthplace, nationality, etc. In untyped network approaches, these distinct fields are aggregated together in some manner to give an overall weight on a link of a single type. Thus two people with the same name have a high link, but so do people born in the same place. This is the approach of Stewart, for example [22], where multiple such kinds of information are merged to yield his single link type.

But in typed networks the source fields can be respected, with each data field represented by a distinct link type. Thus typed link networks can be derived from mutiple relations or tables existing in complex schema. Moreover, they are thus more natural representations for the modern knowledge markup and exchange environments such as XML, RDF(S), and OIL. Since link types also map to predicate types (binary or n-ary if hypergraphs are used), such structures also support inference systems such as used in ontologically-based description logic formalisms (see Fensel *et al.* [8] for a brief discussion of all of this).

3 Typed Link Knowledge Discovery

So assume a large, heterogeneously typed social meta-network structure involving multiple data fields (name, aliases, citizenship, address, age, travel dates, education, etc.) Our goal is then to understand how to generate significant hypotheses concerning such questions as:

1. Which fields are important?
2. For which subsets of the data?
3. Where are the "interesting" areas of structure or activity?

Such questions are obviously closely related to now-classical statistical pattern recognition ("data mining") methods in large multi-dimensional databases, including clustering, feature extraction, classification, etc. [7]. Where our approach differs is its emphasis both on a graph- and network-theoretical representation and a desire to respect the original dimensionalities of the database.

Of course, the most serious problem in moving from untyped to typed structures is the incredible increase in complexity. A network with M nodes is hard enough, with $M(M-1)$ possible binary links. But given in addition N link types, one needs to consider not just all $NM(M-1)$ possible binary links of a *single* type, but moreover, to answer in particular question 1 above, potentially all *combinations* of link types. This number

$$\sum_{n=1}^{N} \binom{N}{n} M(M-1)$$

grows *very* large in N.

Thus our approach is predicated on the idea that fully automatic knowledge discovery methods addressing such questions will *not* be feasible. Instead, we aim at methods which are:

- Appropriate for *moderately* sized multi-graphs (10^2 - 10^5 nodes).
- *Semi-automatic*, and
- *User expert guided*.

The basic idea is to provide an intelligent analyst, a domain expert with background and training in these kinds of mathematical and computer scientific tecniques, with a suite of tools which will support him or her to iteratively guide search for areas of local structure.

4 Link Analysis

It is our intention in future work to provide a complete and rigorous definition of the term "link analysis" in this context as one such broad methodology. This term has a definite, but small, presence in the literature [10]. To our knowledge the concept was developed in the mid 1990's within the law enforcement and anti-money laundering communities (see [19], for example), within which it has considerably more recognition.

It is significant to note that link analysis in our sense of discovery specifically in typed-link networks is usually *not* clearly distinguished from "network analysis" in the sense of single-link networks. An example, again, is Kleinberg [14], whose approach is decidedly network-theoretical in our sense, despite being called link analytical. Thus establishing this term in a proper way may be difficult, but we believe proper to attempt at this time.

The kinds of questions which link analysis is intended to address concern *collections* of records distributed over *collections* of link types. So, for example, given such a collection of records, how do they implicate one collection of link types or another? Similarly, how do they implicate other connnected collections of records, perhaps being more, fewer, or somehow overlapping?

A central concept to our sense of link analysis is known as **chaining**. It works like this:

- Assume a database $\mathcal{D}$ with N dimensions and M data points.
- Define a "view" on $\mathcal{D}$ as its projection to a particular subset of dimensions $n \subseteq \{1, 2, \ldots, N\}$ and restriction to a particular subset of records $m \subseteq \{1, 2, \ldots, M\}$, denoted $\mathcal{D}_{n,m}$.
- Chaining then consists of moving from one particular view $\mathcal{D}_{n,m}$ to another $\mathcal{D}'_{n',m'}$, where $n \cap n' \neq \emptyset, m \cap m' \neq \emptyset$, or both.

Conceptually, first an intelligent analyst considers certain aspects (n) of a certain group of records (m), for example the place of birth (n) of a group of people who all went to the same school (m). She then chains to consider another aspect, say the addresses ($n' \cap n = \emptyset$) of those of that group who went to Harvard ($m' \subseteq m$).

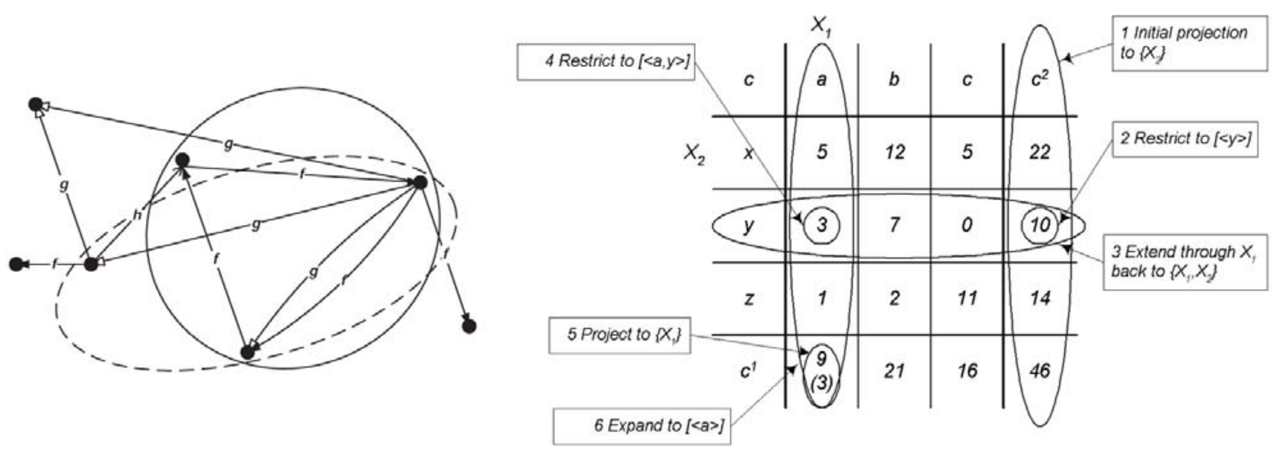

Figure 3: (Left) Chaining in a typed-link network. (Right) Chaining in a two-dimensional contingency table [13].

Fig. 3 illustrates this process in two different contexts. On the left, we have a typed-link meta-network. The solid boundary indicates $\mathcal{D}_{\{f\},\{w,y,z\}}$, the dashed boundary the transition to $\mathcal{D}'_{\{g\},\{x,w,z\}}$, so that $n \cap n' = \emptyset$, but $m \cap m' \neq \emptyset$.

On the right is a somewhat more complex example derived from VizTool, an information theoretical data discovery tool developed at the Los Alamos National Laboratory for fraud detection in IRS

Typed-Link Network	**Abstract Concept**	**MUVE**	**Operationalization**
Nodes	Place	Room	Co-location
Links	Travel	Passageway	Transition in location
Link Type	"Laws of nature"	Who/what travels	Attributes on transitions
		What happens when they do	

Table 1: Relations between typed-link networks and MUVE architectures.

tax databases. VizTool implements our methodology Data Exploration through Extension and Projection (DEEP) [13]. Here the representation is not in terms of labeled graphs, but rather of contingency tables, where the dimensions X_1 and X_2 represent different link types. The cells indicate the number of records with a certain vector value, and the marginal counts are included on the edges of the matrix.

While the complete formal relations between the labeled graph and contingency table representations remain to be detailed in future work, the concept of chaining in both is quite similar. Step 1 indicates an initial view $\mathcal{D}_{\{2\},M}$, all records projected onto the second dimension. The second step restricts this to $\mathcal{D}'_{\{2\},m}$, where $m \subseteq M$ now indicates those ten records $\{\vec{x}\} = \{\langle x_1, x_2 \rangle\}$ such that $x_2 = y$. In the third step, $\mathcal{D}''_{N,m}$ indicates the same set of records, but now extended back both dimensions $N = \{1, 2\} \supseteq \{2\}$. Similar other steps are indicated.

5 Conclusion: Relation to Virtual World Architectures

We conclude with some observations about the potential role of typed-link networks in the architectures of virtual worlds. In particular, we are referring to the Landauer-Bellman sense of a Multi-Uverse Virtual Environment (MUVE): a text-based collaborative environment derived from the Multi-User Dungeon (MUD) heritage [16].

The fundamental metaphor of MUVE spaces is sptial and architectural in the sense of specific places (rooms) and agents which can travel amongst them through their connections (corridors). What's most interesting about MUVEs, however, is the possibility for emergent computational processes which they provide in virtue of the specific properties which exist for rooms, agents, and connections, seen as a "virtual physics" for the artificial space [20].

Space precludes a more detailed discussion, but the mapping into typed-link networks is obvious, and a first cut on this is shown in Tab. 1. Fig. 4 illustrates a cartoon of a MUVE as a typed-link network, where the link types indicate the possible types of connections among rooms.

References

[1] Adamic, Lada A and Huberman, Bernardo: (2000) "The Web's Hidden Order", *Communications of the ACM*, v. **101**, pp. 819-841, Paris

[2] Baron, L; Taguesutclif, J; and Kinnucan, MT et al.: (1996) "Labeled, Typed Links as Cues When Reading Hypertext Documents", *J. American Society for Information Science*, v. **47**:12, pp. 896-908

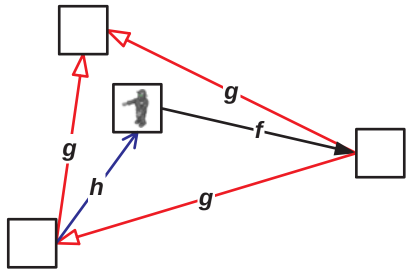

Figure 4: An agent in a MUVE seen as a typed-link network.

[3] Barrett, Chris; Jacob, Riko; and Marathe, Madhav: (2000) "Formal Language Constrained Path Problems", *SIAM J. of Computing*, v. **30**:3, pp. 809-837

[4] Bollobas, Bela: (2001) *Random Graphs, Second Edition*, Cambridge UP

[5] Carley, Kathleen; Lee, Ju-Sung; and Krackh ardt, David: (2002) "Destabilizing Networks", *Connections*, v. **24**:3, pp. 31-34

[6] Davis, Anthony R: (2000) *Types and Constraints for Lexical Semantics and Linking*, Cambridge UP

[7] Fayyad, Usama M; Piatetsky, G; and Smyth, P et al., eds.: (1996) *Advances in Data Mining and Knowledge Discovery*, MIT Press, Menlo Park CA

[8] Fensel, D; Horrocks, J; and van Harmelen, F et al: (2000) "OIL in a Nutshell", in: *2000 Workshop on Applications of Ontologies and Problem-Solving Methodologies*, `http://delicias.dia.fi.upm.es/WORKSHOP/ECAI00/4.pdf`

[9] Gibson, David; Kleinberg, Jon; and Raghavan, Prabhakar: (1998) "Inferring Web Communities from Link Topology", in: *Proc. 9th ACM Conf. on Hypertext and Hypermedia*

[10] Glass, JT; Zaloom, Victor; and Gates, David: (1991) "Computer-Aided Link Analysis (CALA)", *Computers in Industry*, pp. 179-187

[11] Joslyn, Cliff: (1996) "Semantic Webs: A Cyberspatial Representational Form for Cybernetics", in: *Proc. 1996 European Conf. on Cybernetics and Systems Research*, v. **2**, pp. 905-910, `ftp://wwwc3.lanl.gov/pub/users/joslyn/emcsr96.pdf`

[12] Joslyn, Cliff: (2000) "Hypergraph-Based Representations for Portable Knowledge Management Environments", LAUR = 00-5660, `ftp://wwwc3.lanl.gov/pub/users/joslyn/kenv1.pdf`

[13] Joslyn, Cliff and Mniszeiski, Susan: (2002) "DEEP: Data Exploration through Extension and Projection", *Knowledge Discovery and Data Mining*, in preparation

[14] Kleinberg, Jon M: (1999) "Hubs, Authorities, and Communities", *ACM Computing Surveys*, v. **31**:S4, pp. U21-U23

[15] Kleinberg, Jon M: (2000) "Navigation in a Small World", *Nature*, v. **406**

[16] Landauer, Chris and Bellman, Kirstie L: (1998) "Integration and Modeling in MUVEs", in: *Proc. Virtual Worlds and Simulation Conf.*, ed. C. Landauer + K. Bellman, pp. 187-194, Soc. forComputer Sim, San Diego

[17] Newcomb, Steven R; Kipp, Neil A; and Newcomb, Victoria T: (1991) "HyTime: Hypermedia/Time-Based Document Structuring Language", *Communications of the ACM*, v. **34**:11, pp. 67-83

[18] Newman, ME: (2000) "Models of the Small World", *J. Statistical Physics*, v. **101**, pp. 819-841

[19] Office Tech. Asses., : (1995) "Information Information Technologies for Control of Money Laundering", *OTA-ITC-630*, US GPO, Washington DC, `http://www.wws.princeton.edu/~ota/disk1/1995/9529_n.html`

[20] Rocha, Luis M and Joslyn, Cliff: (1998) "Simulations of Evolving Embodied Semiosis: Emergent Semantics in Artificial Environments", in: *Proc. 1998 Conf. on Virtual Worlds in Simulation*, pp. 233-238, Society Comp. Sim., San Diego

[21] Sowa, John F: (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks/Cole, Pacific Grove

[22] Stewart, TA: (2001) *Six Degrees of Mohammed Atta*, in: *Business 2.0*, v. **12/2001**, `http://www.business2.com/articles/mag/0,1640,35253,FF.html`

[23] Watts, Duncan and Strogatz, Steven H: (1998) "Collective Dynamics of 'Small World' Problems", *Nature*, v. **393**, pp. 440-442

[24] Watts, Duncan: (1999) *Small Worlds: The Dynamics of Networks Between Order and Randomness*, Princeton UP