Network Analysis of Toxic Chemicals and Symptoms: Implications for Designing First-Responder Systems

Suresh K. Bhavnani[1] PhD, Annie Abraham[1], Christopher Demeniuk[1], Messeret Gebrekristos[1] Abe Gong[2], Satyendra Nainwal[1], Gautam K. Vallabha[3] PhD, Rudy J. Richardson[4] ScD, DABT
[1]School of Information, [2]School of Public Policy, [4]Toxicology Program, University of Michigan, Ann Arbor, MI; [3]Department of Psychology, Stanford University, Palo Alto, CA

Abstract

The rapid and accurate identification of toxic chemicals is critical for saving lives in emergency situations. However, first-responder systems such as WISER typically require a large number of inputs before a chemical can be identified. To address this problem, we used networks to visualize and analyze the complex relationship between toxic chemicals and their symptoms. The results explain why current approaches require a large number of inputs and help to identify regularities related to the co-occurrence of symptoms. This understanding provides implications for the design of future first-responder systems, with the goal of rapidly identifying toxic chemicals in emergency situations.

Introduction

Toxic chemicals pose a universal threat to humans in situations ranging from bioterrorism to pesticide exposure. In emergency situations such as 9/11, there is a critical need for the rapid and accurate identification of toxic chemicals to reduce harm to large numbers of humans.

To address this need, several organizations have constructed extensive evidence-based databases (e.g., Haz-Map[1] available from the National Library of Medicine [NLM]) that relate toxic chemicals to acute symptoms and properties. Furthermore, there have been attempts to develop devices that make such data accessible to first-responders. For example, NLM has developed the **W**ireless **I**nformation **S**ystem for **E**mergency **R**esponders (WISER[2]), which accepts inputs such as acute symptoms. After each input, the system automatically constructs a database query, and responds with a set of chemicals that satisfy the current set of inputs. As more inputs are received, the set of chemicals narrows to enable the first-responder identify a toxic chemical.

While such systems provide easy access to a database, how effective can they be for pinpointing a toxic chemical in an emergency situation? Toxicologists and public health experts have often reported that acute symptoms and/or properties of toxic chemicals are notoriously non-specific [2,4]. For example, *acute dyspnea* (difficulty breathing) is a symptom caused by a wide range of chemicals (not to mention other health conditions such as *myocardial ischemia* or *asthma).* Therefore, if a first responder enters such non-specific symptoms in WISER, the returned set of chemicals might be too large to be useful. Unfortunately little is known about the overall relationship of toxic chemicals and their symptoms to know whether current approaches are useful or if there might be more powerful ways to assist in the identification of toxic chemicals.

We begin by describing how we estimated the number of symptoms it would take a WISER user to input into the system in order to identify a chemical. We then discuss how we used networks to visualize and analyze the relationship between chemicals and symptoms within the WISER database. The analysis rapidly revealed how symptoms relate to chemicals, and suggested approaches for designing first-responder systems that are better suited to the data. We conclude with a discussion of the network approach for analyzing data, and future research to help improve the rapid identification of toxic chemicals in emergency situations.

Identifying Toxic Chemicals: How Many Symptoms Does it Take?

The WISER system and database is one of the most extensive datasets that relate toxic chemicals to acute symptoms. We therefore analyzed this database to estimate how many symptoms it takes to identify a toxic chemical.

The WISER database (version 2.45) consisted of 390 toxic chemicals, 79 acute symptoms classified in 10 categories (e.g., neurological), and 65 properties classified in 8 categories (e.g., odor). The inclusion of chemicals into the WISER database is determined annually by a team of chemists. Information about each selected chemical is extracted from Hazardous Substances Data Bank (HSDB) (which contains

[1] http://hazmap.nlm.nih.gov/
[2] http://wiser.nlm.nih.gov/

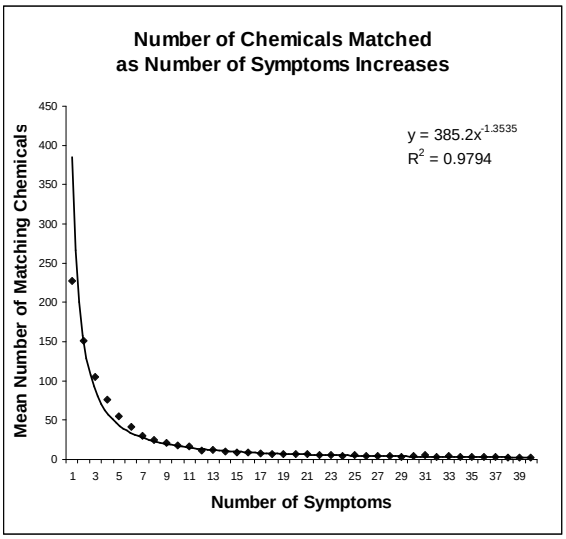

Figure 1. The estimated number of chemicals returned by WISER for different numbers of symptoms follows a power law. The analysis estimates that it takes 40 or more symptoms to uniquely identify a chemical.

extensive information about toxic chemicals targeted to researchers) and inputted into the WISER database (targeted to first-responders).

Because of its target audience and context of intended use, the WISER database neither contains the probability of the association between a chemical and symptom, nor classes of chemicals.

We estimated the number of chemicals returned by WISER for a particular number (*N*) of symptoms as follows. (1) We randomly chose a chemical from the database, randomly selected *N* of its symptoms, and calculated the number of chemicals that matched all *N* symptoms. (2) The above step was repeated 200 times, and used to derive the mean number of matching chemicals for *N* symptoms. Note that the selection of the symptoms, while random, was always based on an existing chemical. This "bootstrap sampling" from the distribution of symptoms captures the co-occurrences of symptoms and closely approximates how symptoms are selected in realistic situations.

Figure 1 shows a plot that estimates the mean number of chemicals returned by WISER based on an increasing number of symptoms. As shown, the analysis estimates that it takes on average about 40 symptoms to uniquely identify a chemical. While this method takes into consideration the co-occurrence of symptoms, it does not take into account real-world variables such as errors in symptom detection and input selection (and therefore is a conservative estimate). One could argue that a first-responder may be able to decide which chemicals are present in the emergency situation by inspecting a small set (e.g., 10) and eliminating most based on contextual information or prior knowledge. However, even in such a scenario it still takes 14 symptoms to narrow the set to 10 chemicals.

Given the large number of symptoms required to identify one (or even 10 chemicals), the current approach of constructing a simple database query to return all relevant chemicals does not appear to be practical for the rapid identification of a toxic chemical in emergencies. This motivated us to analyze why it takes so many symptoms to identify a chemical, and to discover regularities about the relationship between chemicals and symptoms that could be used to develop more effective search methods.

Using Networks to Analyze the Relationship between Toxic Chemicals and Symptoms

Standard statistical techniques such as *distributions* and *cumulative frequencies* collapse data in different ways to provide an overall understanding of the data. However, such techniques are not designed to represent which specific chemicals cause which specific symptoms, therefore potentially concealing important regularities in relationships. To understand such regularities between classes of information, networks are increasingly being used in a wide range of domains [3]. A network is a graph consisting of nodes and edges; nodes represent one or more types of entities (e.g., chemicals or symptoms), and edges between the nodes represent a specific relationship between the entities (e.g., a symptom is caused by a chemical). Figure 2 shows a bi-partite network (where edges exist only between two different types of entities) of toxic chemicals and the symptoms they are known to cause.

Networks have two advantages for analyzing complex relationships. (1) They represent a particular relationship between different nodes and therefore can reveal, for example, regularities in how specific chemicals are connected to specific symptoms. (2) They can be rapidly visualized and analyzed using a toolbox of network algorithms. For example, Figure 2 shows how the *Spring* layout algorithm [1] helps to visualize chemicals and symptoms. The algorithm simulates placing springs between connected nodes, and a weakly repulsive force between nodes that are not connected. As shown, the result is that chemicals that have similar symptoms (e.g., *DDT* and *methoxychlor* in the upper left-hand corner of Figure 2) are placed close to each other, and close to the symptoms that mention them. Given these

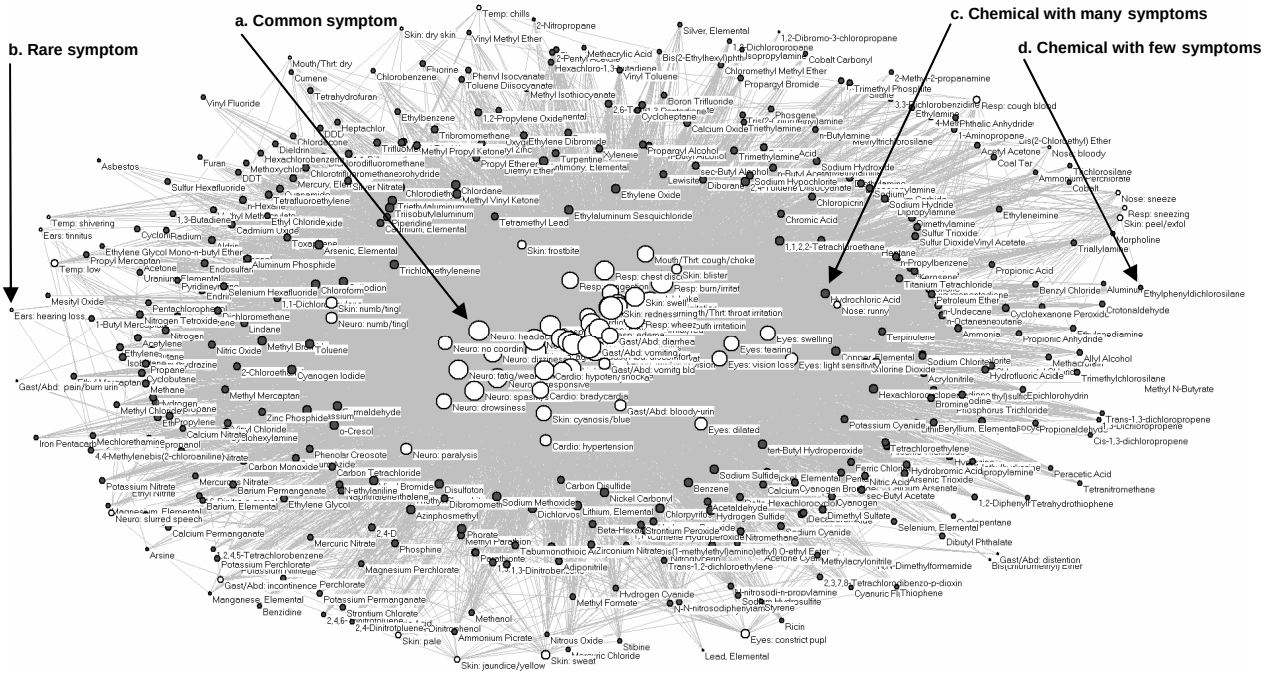

Figure 2. A bi-partite network (automatically generated by the *Spring* algorithm [1]) showing the relationship between 390 chemicals (solid nodes) and 79 symptoms (white nodes). The size of the nodes is proportional to the edges that connect to them. Therefore common symptoms have large nodes, whereas rare symptoms have smaller nodes.

advantages, we explored whether networks could be used to understand the relationship between chemicals and symptoms.

Analysis of the Relationship between Toxic Chemicals and Symptoms

To understand the relationship between chemicals and symptoms, we performed two network analyses: (1) Analysis of the *bi-partite network* shown in Figure 2 to understand why it takes so many symptoms to identify a chemical. (2) Analysis of a *one-mode projection* of the above network to examine the co-occurrence of symptoms. The analyses led to insights for the design of future first-responder systems.

1. Bi-partite network analysis: Why does it take 40 symptoms to identify a chemical?

The bi-partite network shown in Figure 2 was constructed and analyzed using *Pajek* (version 1.17) a network visualization and analysis tool. As discussed earlier, the bi-partite network shows the explicit relationship between the 390 chemicals and the 79 symptoms they cause. Furthermore, besides showing the relationship between the chemicals and symptoms through the connecting edges, the size of a node is proportional to its *degree* (number of edges that connect to that node). Therefore, the larger a node, the more edges it shares with other nodes.

The bi-partite network visualization revealed two critical patterns related to symptoms and chemicals.

(1) There are 59 commonly-occurring symptoms in the center of the network, while 20 rare symptoms are placed around the periphery. For example, *Neuro: headache* (a) is in the center of the graph with 257 edges each connected to a chemical. In contrast *Ear: hearing loss* (b) is on the far left periphery with only 7 edges. Therefore, the mean degree of symptoms is large, and there is a large range in degrees (Mean=154.73, SD=106.97).

(2) The chemicals form a ring around the 59 symptoms in the center. Chemicals close to the inner set of symptoms cause many symptoms compared to chemicals in the outer ring. For example, the chemical *Hydrochloric Acid* (c) has 47 symptoms, whereas the chemical *Ethylphenyldichlorosilane* (d) has 17 symptoms. Therefore, the mean degree of chemicals is small, and there is a small range in degrees when compared to symptoms. (Mean=31.34, SD=10.53).

The above network structure in which there are many chemicals (in the ring) with similar degree, and a relatively smaller number of symptoms (in the center) with high degree, results in a high overlap in the number of symptoms for most chemicals. This can be seen in the high density of edges[3] (resulting in a gray

[3] Edge density (number of actual edges / number of possible edges) = 0.396. This is high compared to the edge density of most large networks that have been analyzed, which typically ranges from almost zero to 0.1 [3]. (Edge density for a fully connected network = 1.0).

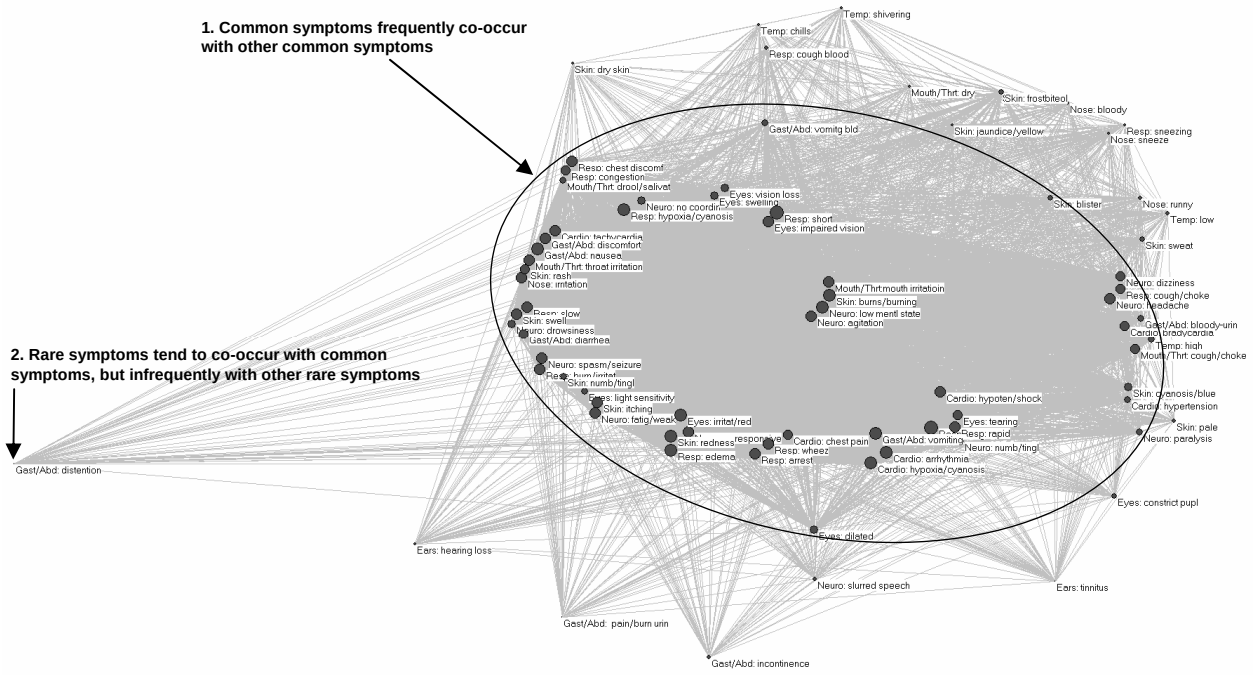

Figure 3. A one-mode projection of the network in Figure 2 showing regularities in the co-occurrence of 79 symptoms.

mass of indistinguishable edges) between the chemicals in the ring, and the symptoms in the center.

The above result provides an explanation for the power law curve in Figure 1. The curve has its shape because after the first few symptoms have helped to rapidly distinguish between chemicals that have widely different symptoms (leading to the steep initial drop), there are fewer and fewer symptoms left to discriminate between many chemicals with highly overlapping symptoms (leading to the long tail). If however, the overlap between chemicals was not high, one could expect the curve to drop rapidly, but have a shorter tail.

While the bi-partite graph revealed why it takes so many symptoms to identify a chemical, it concealed the specific co-occurrence pattern between the symptoms. To reveal the co-occurrence between symptoms, we constructed a *one-mode projection* of the bi-partite network to analyze patterns in the co-occurrence of symptoms.

2. One-mode projection analysis: How do symptoms co-occur?

As large networks with many edges can get visually complex, there exist many methods to transform the data in order to uncover hidden relationships. One such method is the *one-mode projection* of a network, which in our case removes all chemical nodes and adds edges between symptom nodes that are caused by the same chemical. For example, if the chemical *DDT* causes *edema* and *vomiting*, then the one-mode projection will remove the *DDT* node, and add an edge between the *edema* and *vomiting* nodes. Furthermore, if these two symptoms are connected to another chemical, the weight of the edge between them increases. The edge weight therefore represents how many times a pair of symptoms co-occurs. The resulting *Spring* layout places symptom nodes close to each other if they have high co-occurrence, and spaces them far apart if they have low co-occurrence.

Figure 3 shows a one-mode projection of the network in Figure 2. The network reveals two regularities. (1) There is a densely connected set of high degree symptom nodes in the center that frequently co-occur with each other (the edges are so dense that they cannot be seen individually). This set contains 57 of the 59 symptoms that were in the core of Figure 2. (2) Rare symptoms in the periphery of the network tend to co-occur with the common symptoms in the center, but infrequently with other rare symptoms[4].

Discussion

The bi-partite network and the one-mode projection together provided answers for two questions. (1) Why *does it take 40 symptoms to identify a chemical?* The analysis revealed that this is because of the high overlap of symptoms between most chemicals. (2) *How do symptoms co-occur?* The analysis revealed

[4] The network therefore exhibits almost no *degree correlation* [3] (Pearson's correlation = 0.008) between pairs of nodes, when taking into account the edge weights.

that a core set of common symptoms are densely connected to each other, and rare symptoms co-occur with common symptoms but infrequently with each other.

The above observations do not appear to be unique to WISER. Our preliminary analysis of symptoms and health conditions in other datasets (e.g., Collaborative on Health and the Environment Toxicant & Disease Database), suggest that they have network properties similar to WISER. We therefore explored implications of our network analysis for the design of future systems to help search such datasets.

Implications for Designing First-Responder Systems

The results of our network analyses suggested a *multi-input* approach to help first-responders to rapidly identify chemicals. This approach will provide three different ways to select a symptom:

1. Select from a static set of symptoms presented in a hierarchy (as is currently provided by WISER). This approach is suitable if the user knows the exact name of a symptom, and can rapidly identify its location in the hierarchy. However, as our analysis has shown, selecting symptoms which are then converted into a database query is not the most efficient method to identify a chemical.

2. Select from a dynamically generated list of symptoms (using a dynamic binary search tree [BST] algorithm) ranked by the ability of a symptom to eliminate close to half of the remaining chemicals. This approach is suitable if the user wishes to rapidly narrow the set of chemicals, based on suggestions from the system. Our initial experiments using a BST suggest that it can substantially reduce the number of symptoms to identify a chemical. This might be because of the confluence of network properties such as low degree correlation and high edge density, a hypothesis that needs to be tested in future research.

3. Select from a dynamically generated list of symptoms ranked by co-occurrence of already selected symptoms. This approach can be used to check which symptoms should be co-occurring with the ones that have been selected, and was suggested by the co-occurrence patterns of symptoms. Selection from a list of co-occurring symptoms is not designed to reduce the number of symptoms, but rather to provide useful feedback for the first-responder.

Future research should determine whether the above multi-input approach (each with different trade-offs) provides improvements over the current WISER approach, and if it can be useful for other datasets.

Summary and Future Research

Given the critical importance of rapidly identifying toxic chemicals by first-responders, we investigated the relationship of chemicals and their symptoms in two steps. First, we analyzed how many symptoms on average it would take to identify a unique chemical using the current WISER system developed by NLM. The analysis revealed a conservative estimate of 40 or more chemicals. Because such a high number appeared impractical in emergencies, in our second step, we analyzed the relationship between chemicals and symptoms using networks. The analysis revealed a high overlap in the symptoms between chemicals, and co-occurrence patterns in the symptoms. These results led to insights about how to design future systems that could help first-responders rapidly identify toxic chemicals.

The network analysis tools we have demonstrated are only a small subset of those available. In our future research, we will apply a broader range of network analysis methods to WISER and other public health databases. Furthermore, the WISER database has other chemical properties, such as color and odor, which no doubt figure in the identification of chemicals and therefore will be a focus of our attention in future analysis and design. The use of networks, as demonstrated in this preliminary study, should therefore lead to new regularities about the data that can be exploited, with the goal of helping in the rapid identification of chemicals and other determinants of acute and chronic symptoms.

Acknowledgements

We thank Bernstam, E., Ganesan, A., Jansenn, A., Lechman, D., Schettler, T., and Varaprath. S. for their feedback and assistance in interpreting the WISER data, and Adamic, L. for teaching us how to analyze networks. Our sincere thanks to Bijan Mashayekhi and Philip Wexler from NLM who graciously provided us access to the WISER database.

References

1. Freeman, L. Visualizing Social Networks, *JoSS*, 1,1 (2001).
2. MMWR. Recognition of illness associated with exposure to chemical agents—United States. *Morb. Mortal. Wkly. Rep*. 52, (2003) 938-940.
3. Newman, M. The structure and function of complex networks. *SIAM Review*, 45(2), (2003), 167-256.
4. Schier, J.G., Rogers, H.S., Patel, M.M., Rubin, C.A., Belson, M.G. Strategies for recognizing acute chemical-associated foodborne illness. *Military Medicine* 171, (2006) 1174-1180.