Path and Relationship discovery using Sparse Recovery and Compressive Sensing

Harshal Chaudhari
Boston University
harshal@bu.edu

Yu (Albert) Chen
Boston University
chenyua@bu.edu

Shan Sikdar
Boston University
ssikdar1@bu.edu

Jacqueline You
Boston University
jgyou@bu.edu

Abstract

Frequently real world problems involve datasets with large number of instances and small partial information about the relationship between individual instances. Often it is important to infer the structure of the data so that given a data point one can discover its relationship to other nodes to discover important information. In this paper we have attempted to solve this problem by reducing the problem to path discovery. Our dataset was obtained from the Identity Discovery Challenge. Given a partial New York city phone number and an address in Maryland, we were tasked with finding the relationship between the two pieces of information in order to identify a male who could potentially be carrying a contagious disease. A variety of compressive sensing and sparse methods were utilized in order to solve the challenge including sparse graph recovery, shortest path algorithms, page rank, and compressive sensing techniques such as L1 Minimization. Our experiments suggest that relationships and paths can be inferred and used in real life applications for identity discovery.

1 Introduction

Compressive sensing has emerged within the past decade as a new method of signal/image sampling that challenges traditional methods that rely on high sampling rates. This technique was first described in works [Candes, Donoho etc.] and centers on two concepts - sparsity and incoherence. Sparsity expresses the idea that the information of a signal may much be smaller than suggested by its bandwidth. Incoherence describes the modality of the data by describing that signals with sparse representations must be spread out in the domain that they are acquired in. [Candes, Wakin].

Compressive sensing's applications extend beyond digital images and digital signals. Compressive sensing can also be utilized to examine cliques and groups. [Shi,Tang,Xu, Moscibroda] Compressive sensing has also been applied in a variety of ways in medicine, including increasing the efficiency of magnetic resonance imaging (MRI) through compressive sensing reconstruction [Lutsig et al]; using compressive sensing approaches (Gradient Projection for Sparse Reconstruction and Belief Propogation) for screening of rare genetic diseases [Erlich etal]; and by using probabilistic tests and compressive sensing to identify a group of infected individuals [Cheraghchi et al].

In our work, we attempted to use compressive sensing to tackle a hypothetical epidemiological scenario. We used compressive sensing to aid us in path discovery in a graph. We experimented and attempted many methods including sparse graph recovery, L1 minimization, page rank and shortest path algoirithms.

2 Problem Premise

We were tasked with determining the identity of an unknown male individual given a fictitious set of records. In this hypothetical scenario, the unidentified man had visited a Maryland hospital and had

potentially infected another patient with a deadly disease. Authorities trying to locate this man only had an incomplete version of his phone number (with a New York City area code). The challenge was to identify the individual using these facts and a data set of records.

The data set consisted of a list of approximately 350,000 nodes and another list of approximately 68,000 edges corresponding to these nodes. Each node consisted of the following attributes: first name, last name, middle name, street, city, state, zip, phone number, and ID-DOC. The man's incomplete phone number `21299875XX` was designated as "Seed 1" while the hospital's address, `4408 East Madison Ave., Bethesda, MD 20014`, was designated as "Seed 2."

The edges were produced through an entity resolution and fuzzy matching process, and for each pair of edges, there was a set of attribute pair scores (APS) which were then combined to yield a total composite score (TCS). When represented as an adjacency matrix, this data set was sparse due to the relatively low number of edges.

Sample Node

Source	GUID	LastName	MiddleName	FirstName	Street	City	State	Zip	Phone	ID-DOC
CCCR	CCCR-a57685ee-ba9f...	Brandybuck	Donnamira	Gloriana	2719 Pin Oak Drive	Manhattan	NY	10018		5334856597493120

Sample Edge

EDGE_ID	GUID_1	GUID_2	Last_APS	Mid_APS	First_APS	Street_APS	City_APS	State_APS	Zip_APS	Phone_APS	ID-DOC_APS	TCS
Edge_2421	SRLU-2a...	CCCR-a...	1	0.7777778	1	0.05263	0	1	0	0	0	0.4788

We decided that this problem eventually boiled down to path discovery between the two SEED nodes. We then began exploring and experimenting with ways to most efficiently determine this path.

3 Linear Regression

We first attempted to recover more edge information of the graph by using linear regression to produce an equation that could help derive us more edges between nodes of the graph. Since the edge contained both an "Attribute Pair Score" (APS) for each variable and a "Total Composite Score"(TCS) for each i, we believed that the total composite score (TCS) for each ith pair of nodes could be expressed as follows:

$$TCS = b_0x_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$$

where x_k is an APS attribute score for a particular variable. Using linear regression intended to determine the formula's constants used to calculate the TCS score. This would allow us to then for any two nodes utilize fuzzy matching to create APS scores and use the equation to calculate a TCS to be the edge wieght.

However, we discovered that the runtime for reconstructing the graph would be $O(n^2)$; thus this approach was not suitable for our data.

4 Page Rank Algorithm

PageRank is an algorithm developed by Larry Page from Google used to rank websites in their search engine. It is a link analysis algorithm that tries to weight each element in order to measure it's realtive importance to the entre collection. The algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on link will arrive at any particular page. We attempted to use PageRank to solve this challenge by using the information stored in the nodes as opposed to web content.

5 Finding Shortest Path

We proceeded to treat the data set as a graph $G = (V, E)$, where V corresponds to the individual records/nodes and E corresponds to the linkages between records with an edge weight of $(1-TCS)$.

We sought paths between Seed 1 and Seed 2 with total edge weights below a designated threshold so that we could identify a number of potential disease carriers along with related individuals (family, friends, et cetera). We then used the provided information and the data set to determine which individuals were most likely to be our person of interest. As detailed in our results, this method worked successfully; however, it suffered from a lack of generalizability to other problems. [or add some other explanation]

6 Compressive Sensing Techniques

6.1 Radon Basis Pursuit

[Jiang et al] introduced a network data representation framework that allowed for the use of compressive sensing to study the network clique detection problem. They demonstrated that higher-level cliques could be extracted from analysis of a group of people by using radon basis pursuit. However, a number of limitations to this approach include the combinatorial increase in basis size, and the need for a binary weight of edges (i.e. edges are weighed either 0 or 1). These limitations meant that this approach ultimately was not suitable for our problem.

6.2 Sparse Recovery with Graphs

Previous work in sparse recovery of graphs has been done in situations in which a weight is associated with each vertex. [Wang,Xu, Mallada, Tang] . Using this approach we hoped to obtain communities of nodes similar to the original SEED node.

Consider the graph G = (V,E), V the set of all verticies $x_1, x_2, ..., x_n$ and E the set of edges between any two vertices. Consider only the connected components of G. Let $x = \{x_1, x_2, x_3, ..., x_n\}$. Note that x should be sparse since we assume very few nodes to be similar to the SEED-1 node. Let the support of the vector x be defined to be a vector with its entries being the non-zero indicies of x, in other words: $\text{supp}_x\text{`} := \{x_i | x_i \neq 0\}$. Then, $||x||_0 = |\text{supp}_x|$. x is defined to be k-sparse if $|\text{supp}_x| = k$.

Look at a subset of nodes $S \subseteq G$. Define E_S as the subset of edges with its vertices in S. Then $G_S = (S, E_S)$ is a induced subgraph of G. From here two assumptions must be made about measurement of S. First a set of nodes S can be measured iff G_S is a connected subgraph. Secondly the measurement of S must be an additive sum of values at the corresponding nodes. Let the measurement is defined to be an additive sum of all real values of nodes. A single specific measurement is obtained by randomly selecting a node in the connected compontent and performing an additive sum of the tree with depth 'h' from the given node. In relation to the Identity challenge, this measurement would represent how similar a particular induced subgraph of G is similar to the SEED node.

Let $y = \{y_1, y_2, y_3 ... y_m\}$ denote randomly chosen measurements such that $m << n$. Let A be an mxn measurement matrix such that its ith row corresponds to the i measurement. For a particular entry A_{ij}, $A_{ij} = 1$ if node j is included in the i'th measurement and $A_{ij} = 0$ otherwise. Note that we can write the above situation much more compactly as $y = Ax$. This equation has now been written in the same from used in compressive sensing and sparse recovery. Compressive sensing theory suggests that the n-dimensional vectors can be found from m measurements if the vectors are sparse enough. Therefore l_1 minimization can be used to recover the sparse vectors using 'm' ramdom measurements. After the vectors are recovered, the maximal entries in x will deonte the nodes most similar to the SEED node. Thus this can be seen as the set of potential disease carriers and a shortest path approach should be able to identify the person.

6.3 L_1-minimization

should this be here? or should a discussion about which L1 minimization technique used from L1 magic in the section above and/or results

7 Results

7.1 Linear Regression

Using linear regression method described above, we successfully found the relevant weights. The results endedup being very interesting with the APS of ID-DOC having that largest weight. Upon browsing through the dataset it was shown the the ID-DOC did not appear often compared to other variables. A high weight would possibly indicate that when two nodes have a close matching ID-DOC score that this is a very important relationship.

$APS_{LastName}$	$APS_{FirstName}$	APS_{Middle}	APS_{phone}	APS_{state}	APS_{City}	APS_{Street}	APS_{Zip}	APS_{ID-DOC}
-0.1293	-0.1708	-0.3292	0.1341	-0.5251	0.1580	0.2225	-0.1707	1.0561

7.2 Page Rank

For page rank we hoped that the rank of Seed-1 and Seed-2 would indicate which other nodes of information was most closley relevant, however the rank of Seed-1 and Seed-2 turned out to be average so it was impossible to draw out any further information.

7.3 Shortest Path

The initial shortest path returned a series of nodes that primarily included information about a woman. Since the information provided indicated the person of interest was male, the last edge between the final pair of nodes on the shortest path was removed and the shortest path algorithm was run again. This time the shortest path eventually led us to nodes containing information about a male. Upon investigating into the intersection sets of the path information was found possibly sugessting that the woman in the first path was in fact to the man indentified in the second path. After obtaining the second shortest path we again removed the last edge in the path and ran shortest path on the remaining graph.The algorithm yieled no shortest path between the two nodes.

First Shortest Path:

Source	LastName	MiddleName	FirstName	Street	City	State	Zip	Phone	ID-DOC
Source	LastName	MiddleName	FirstName	Street	City	State	Zip	Phone	ID-DOC
SEED	NaN	NaN	NaN	NaN	NaN	NaN	NaN	21299875XX	NaN
SRLU	Brandybuck	D	Gloriana	3306 Rosewood Lane	New York	NY	10003	2129987506	NaN
CCCR	Brandybuck	Donnamira	Gloriana	2719 Pin Oak Drive	Manhattan	NY	10018	NaN	5.33E+15
CCTR	NaN	NaN	NaN	18 Wayback Road	Bethesda	MD	20014	NaN	5.33E+15
CCTR	NaN	NaN	NaN	1323 Frosty Lane	Lodi	NY	14860	NaN	4.49E+15
SEED	NaN	NaN	NaN	4408 East Madison Ave.	Bethesda	MD	20014	NaN	NaN

Second Shortest Path:

Source	LastName	MiddleName	FirstName	Street	City	State	Zip	Phone	ID-DOC
Source	LastName	MiddleName	FirstName	Street	City	State	Zip	Phone	ID-DOC
SEED	NaN	NaN	NaN	NaN	NaN	NaN	NaN	21299875XX	NaN
SRLU	Brandybuck	D	Gloriana	3306 Rosewood Lane	New York	NY	10003	2129987506	NaN
CCCR	Brandybuck	Donnamira	Gloriana	2719 Pin Oak Drive	Manhattan	NY	10018	NaN	5.334857e+15
HPA	Took	NaN	Tollman	Pin Oak Dr	Manhattan	NY	10018	NaN	NaN
ID	Took	Fredegar	Tolman	234 Trails End Rd.	Staten Island	NY	10301	NaN	298808448
TR	Tuk	F	Tom	NaN	NaN	NaN	NaN	6318085343	298808448
HR	NaN	NaN	NaN	322 Meadow Dr.	Bethesda	MD	20014	6318085343	NaN
WP	Hornblower	NaN	Melilot	322 Doe Meadow Drive	Bethesda	MD	20014	3018035414	NaN
SEED	NaN	NaN	NaN	4408 East Madison Ave.	Bethesda	MD	20014	NaN	NaN

7.4 Sparse Recovery of Graph

8 Conclusion

References

[A0] CANDÉS, E., ROMBERG, J., AND TAO, T. 2006. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEE Trans. Inform. Theory 52*, 2, 489-509.

[A1] CANDÉS, E. AND TAO, T. 2006. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory 52*, 12, 5406-5425.

[A2]

[A3] CANDÉS, E. J. AND WAKIN, M. B. 2008. An introduction to compressive sampling. *IEEE Signal Processing Magazine, 25*, 2, 21-30.

[B] LUTSIG, M., DONOHO, D., AND PAULY, J. M. 2007. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine, 58*, 6, 1182-1195.

[C] ERLICH, Y., SHENTAL, N., AMIR, A., AND ZUK O. 2009. Compressed sensing approach for high throughput carrier screen. In *47th annual Allerton conference on communication, control, Monticello, IL, IEEE, 539-544.*

[D] CHERAGHCHI, M., HORMATI, A., KARBASI, A., AND VETTERLI, M. Group testing with probabilistic tests: Theory, design and application.

[X] JIANG, X. Y., YAO, Y., LIU, H., AND GUIBAS, L. 2012. Detecting network cliques with radon basis pursuit. In Proceedings of the 15th international conference on artificial intelligence and statistics (AISTATS), La Palma, Canary Islands, Spain, April 2012, Journal of Machine Learning, 565-573.

[?] Wang, Xu, Mallada, Tang. Sparse Recovery with Graph Constraints: Fundamental Limits and Measurement Construction.

[?]SHI T., Tang D., Xu L., Moscibroda T. In Correlated Compressive Sensing for Networked Data. The 30th Conference on Uncertainty in Artificial Intelligence (UAI)