Inferred Networks and the Social Determinants of Health

Prashant Sanjel and John Matta

Southern Illinois University Edwardsville, Edwardsville, IL 62026, USA,
psanjel,jmatta@siue.edu

Abstract. This paper explores the social determinants of health through a network science based approach to analyzing the Latino MSM Community Involvement (LMSM-CI) dataset. Data are clustered to determine identifying characteristics of groups of participants in 3 categories: high self esteem, susceptibility to alcohol abuse, and HIV positive status. A question arises as to the best methodology for inferring a graph from the data, as well as for clustering and analyzing the network. To that end we use 4 different graph inference methods: inverse covariance selection (Glasso), neighborhood selection (MB), Sparse Correlations for Compositional data (SparCC) and the traditional k-Nearest Neighbors (kNN). For each inference we test 4 different clustering methods: Louvain, Leiden, NBR-Clust with VAT, and NBR-Clust with integrity. Surprisingly, the Glasso and MB inference methods produce better clusterings than kNN, as determined by a suite of internal evaluation measures. The most promising clusterings are visualized and their properties are analyzed.

Keywords: graph inference, clustering, machine learning, health data analysis

1 Introduction

The desire to find a happy and healthy life has led to the emergence of Social Determinants of Health (SDoH). The Centers for Disease Control (CDC) outlines SDoH as "conditions in the environments where people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks" [10]. Those conditions can be a catalyst for better health and prosperity. Understanding data on SDoH such as education, income, employment, age, and more can aid in attempts to improve community health. The use of machine learning algorithms such as clustering on graphs inferred from health and medical datasets presents opportunities towards personalized medicine, and a healthier and better quality of life.

There are benefits to inferring a network from existing data and working with it in network form [7]. These include the ability to simplify data to a nodes-and-edges graph representation, the opportunity to employ a wide range of existing algorithms and tools, and the ability to learn from effective visualization of the information. Each of these benefits is put to use in this study.

Due to technical advances in genetic sequencing technologies there has been a large influx of new data. Desire to exploit this data has lead to a variety of new graph inference techniques, including those examined here: inverse covariance selection (Glasso) [14], Meinshausen and Bühlmann (MB) neighborhood selection [23], Sparse Correlations for Compositional data (SparCC) [13], and the traditional k-Nearest Neighbors (kNN). Due to expanded use of electronic medical records, there is a large amount of newly-available medical data. This paper seeks to determine the applicability of these graph inference methods to applications with health and medical data. Surprisingly, Glasso and MB give strong results, which is a contrast with previous papers that preferred kNN graphs [17].

Social determinant data used here are from Jesus Ramirez-Valles' *Latino MSM Community Involvement: HIV Protective Effects* (LMSM-CI) survey [27], the original purpose of which was to determine whether community involvement reduced the risk of HIV in male Latino populations. We examine social determinants leading to target variables of high self esteem, alcohol abuse risk, and positive HIV status via graph inference methods and clustering on the LMSM-CI dataset. The data are available online through application with the National Addiction & HIV Data Archive Program [1]. This research was approved by the SIUE Institutional Review Board.

2 Related Work

In this study we have inferred Glasso, MB, SparCC, and kNN graphs. The Glasso method is based on graphical lasso regression, which has been widely used in diverse applications like financial interpretation [4] and facial recognition [32]. KNN is a popular network inference method, with previous medical data applications including predicting sleep disorders [26] and analyzing stroke biomarkers [11]. In [18] the graph methods that we see in this paper are tested with microbial associations. They have also have been applied to other domains, such as metagenomics and inflammatory bowel disease [1], ribosomal and micro RNA [13], and spoilage of food [25].

Machine learning approaches have been employed in the study of social determinants of health [16], such as in [28], where social and economic factors are used to predict medical data like blood pressure and body mass index. In other examples, social and economic factors have been used to study the causes of addiction [2] and to evaluate mental disorders [3].

This paper focuses on four network-based methods of community detection: Louvain, Leiden, and NBR-Clust with VAT and integrity. The best clusterings are chosen by cluster evaluation methods, in accordance with work previously done in [17]. Machine learning tools like clustering have been used on graphs inferred from medical data to analyze and understand conditions such as autism spectrum disorder [22]. Clustering has been used to predict drug-drug reactions [30] and analyze DNA sequences [15]. In this study we rigorously analyze graph inference and machine learning techniques as applied to the LMSM-CI dataset.

[1] https://www.icpsr.umich.edu/web/pages/NAHDAP/index.html

3 Methods

3.1 The LMSM-CI Dataset

The data used in this study are from the Latino MSM Community Involvement: HIV Protective Effects survey (LMSM-CI) [27]. The data were collected in 2003-2004 and consist of 323 samples of Latino men who have sex with men (MSM) in the Chicago, Illinois, USA metro area and 320 samples in the San Francisco, California, USA area. The study consists of over 900 variables, with many variables related to the social determinants of HIV in the Latino MSM community. The foremost aim of the survey is to determine if Latino MSMs are more likely to take precautions against HIV if they are involved in community activities. The large amount of data collected makes this dataset relevant not just to HIV, but to other determinants as well.

3.2 Data Curation

Many of the variables contained in the LMSM-CI dataset were meta-data, such as the IDs of the recruiter and subsequent recruitees of the sample. Also, the survey contained many questions involving participation in specific charities or social groups. To simplify analysis, these variables were removed. The remaining multi-valued variables were converted to binary variables using one-hot encoding. As an example of one-hot encoding, a variable such as *employment-status* with possible answers *employed*, *unemployed*, and *on-disability*, is changed to three variables with yes/no answers: *employment-status:employed*, *employment-status:unemployed*, and *employment-status:on-disability*.

The National Institute on Alcohol Abuse and Alcoholism (NIAAA) [24] defines heavy alcohol abuse risk for men as binge drinking for more than 4 days in the past month, where binge drinking is defined by the Substance Abuse and Mental Health Services Administration (SAMHSA) as drinking more than 5 alcoholic drinks. We created the target variable *Alcohol Risk* based on these definitions and the LMSM-CI variables concerning frequency of alcohol consumption and number of drinks on a typical day. The self esteem target we study represents an affirmative answer in the survey to “I have a positive attitude about myself.”

Feature selection was performed as described in [11], with stratified testing and training sets. The step forward algorithm was used to select the features giving the best performance as determined by logistic regression. The end result was a list of the most important 70, 20 and 15 variables for each of the 3 target variables being studied.

3.3 Graph Inference

The LMSM-CI data were converted into graph format using four graph inference methods. These inference methods assume sparse networks, so in all cases parameter settings were adjusted to produce a connected graph with the minimum number of edges. The first graph inference method is the widely used k-Nearest Neighbors (kNN), where distances are calculated between each pair of proband vectors $\overrightarrow{uv}$, and an edge is placed between node u and its k shortest distance neighbors v. The kNN graphs were created using the CCCD R package [19].

The second graph inference method is Glasso. This is a fast method for "estimating sparse graphs by a lasso penalty applied to the inverse covariance matrix" [14]. The third inference method, Meinshausen and Bühlmann (MB) [23], uses neighborhood selection. It "estimates a sparse graphical model by fitting a lasso model to each variable, using the others as predictors" [14]. The last method is Sparse Correlations for Compositional data (SparCC) [13]. SparCC uses linear Pearson correlations to infer a network of associations. This approach has been shown to ameliorate unreliable results that can occur with correlation analysis methods. Glasso, MB, and SparCC networks were created using Sparse Inverse Covariance Estimation for Ecological Association Inference (SPIEC-EASI) [18].

3.4 Clustering

Four clustering methods were used. Louvain [6] is a popular, low-time-complexity algorithm that seeks to produce clusters that maximize modularity, a well-known method of quantifying the goodness of a clustering, as measured against a random clustering. The Leiden algorithm is meant to represent an improvement over Louvain. It "converges to a partition in which all subsets of all communities are locally optimally assigned. The Leiden algorithm is faster than the Louvain algorithm and uncovers better partitions. In addition, it has been proved that the Leiden algorithm yields communities that are guaranteed to be connected" [29]. The NBR-Clust framework [21] uses network resilience measures to partition a graph into clusters. It identifies an attack set of nodes $S \in V$, whose removal partitions the network into some number of disconnected components. Resilence measures used with NBR-Clust are integrity, which is defined as

$$I(G) = \min_{S \subset V} \left\{ |S| + C_{max}(V - S) \right\}, \tag{1}$$

and vertex attack tolerance (VAT), which is defined as

$$VAT(G) = \min_{S \subset V} \left\{ \frac{|S|}{|V - S - C_{max}(V - S)| + 1} \right\}, \tag{2}$$

where V is the set of vertices, S is the attack set, and C_{max} is the size of the remaining largest connected component. Resilience measures are approximated using betweenness centrality [20]. For all three clustering methods, the number of clusters is not specified *a priori*.

3.5 Cluster Evaluation

We quantified the success of clustering results with 5 internal evaluation measures, using the methodology presented in [22, 17]. Cluster evaluation measures

Table 1. Cluster Evaluation Results for Self Esteem

Cluster Algorithm	#	**Glasso**					**SparCC**				
		Davies-Bouldin	Silhou-ette	Calinski-Harabasz	Baker-Hubert	Hubert-Levine	Davies-Bouldin	Silhou-ette	Calinski-Harabasz	Baker-Hubert	Hubert-Levine
Louvain	15	1.911	0.168	4.376	0.546	0.248	2.216	-0.021	1.589	0.240	0.396
	20	1.558	0.199	5.593	0.708	0.155	2.399	0.033	1.834	0.223	0.403
	70	2.354	0.054	2.293	0.426	0.330	2.275	2.275	1.841	0.297	0.384
Leiden	15	2.402	0.174	4.416	0.650	0.209	2.137	0.031	2.014	0.262	0.369
	20	1.558	0.199	5.593	0.708	0.155	2.387	0.028	1.820	0.215	0.402
	70	2.648	0.093	2.648	0.526	0.281	2.426	0.042	2.038	0.321	0.372
VAT	15	1.974	0.096	3.816	0.573	0.239	1.711	-0.084	1.231	0.008	0.499
	20	2.130	0.075	4.907	0.458	0.262	1.718	-0.210	0.981	-0.105	0.563
	70	2.351	-0.164	1.511	0.180	0.429	1.243	-0.226	0.822	-0.182	0.550
Integrity	15	1.974	0.096	3.816	0.573	0.239	1.794	-0.131	1.194	0.082	0.467
	20	1.263	0.003	2.786	0.577	0.228	1.718	-0.210	0.981	-0.105	0.563
	70	2.086	-0.174	1.261	0.208	0.449	1.598	-0.226	1.004	0.121	0.476

quantify (in different ways) desirable clustering properties, such as maximal separation between clusters combined with minimal separation within a cluster. The following evaluation methods were used: Davies-Bouldin, Silhouette, Calinski & Harabasz, Baker & Hubert, and Hubert & Levine, from the ClusterSim R package [31]. A higher score indicates a better clustering, except with Davies-Bouldin and Hubert & Levine, where a lower score indicates a better clustering.

Each clustering result was given a score, determined by an ensemble method, where the clustering with the best score for each evaluation measure was given a point, and the clustering with the most points was chosen. Ties were broken by comparing individual scores between the tied instances. For example, between two tied clusterings, the clustering with a better score on an evaluation method such as Davies-Bouldain would be assigned 1 point. This would be repeated for all 5 evaluation methods, and the highest-scoring clustering would be chosen.

4 Results

4.1 Cluster Evaluation

Results from cluster evaluation are shown in Tables 1, 2, and 3. Each line in a table represents a clustering method in combination with the number of variables used to infer the graph (15, 20, or 70). Highlighted numbers represent the best scores for each clustering algorithm and receive 1 point. Total points determine the best clustering. Due to space limitations, results for some graph types are not shown. For self esteem, the high scoring choice was Glasso-Leiden-20, as shown in Table 1. For alcohol risk, as shown in Table 2, both Glasso and MB clusterings performed well, scoring highest on 3 evaluation measures. The tie between Glasso-Leiden-15 and MB-VAT-20 was broken in favor of MB-VAT-20. For HIV results, the Glasso and MB clusterings performed much better than SparCC and kNN, and are shown in Table 3. The 3-way tie was broken in favor of Glasso-VAT-15. Based on these results, we visualize and analyze the Glasso-

Table 2. Cluster Evaluation Results for Alcohol Risk

Cluster Algorithm	#	Glasso					MB				
		Davies-Bouldin	Silhouette	Calinski-Harabasz	Baker-Hubert	Hubert-Levine	Davies-Bouldin	Silhouette	Calinski-Harabasz	Baker-Hubert	Hubert-Levine
Louvain	15	1.624	0.154	10.203	0.700	0.164	1.867	0.011	4.919	0.484	0.272
	20	2.623	0.144	8.020	0.590	0.226	1.813	0.171	11.245	0.643	0.187
	70	3.282	0.024	3.365	0.395	0.338	2.612	0.017	2.151	0.250	0.412
Leiden	15	1.639	0.113	10.475	0.611	0.216	1.893	0.096 -		0.502	0.277
	20	2.175	0.117	8.280	0.518	0.253	2.238	0.120	4.967	0.513	0.265
	70	3.088	-0.030	3.064	0.348	0.369	2.409	0.027	2.282	0.279	0.397
VAT	15	1.659	-0.017	6.069	0.588	0.218	1.807	-0.013	3.898	0.327	0.341
	20	2.271	-0.125	4.348	0.332	0.337	1.364	0.190	10.401	0.627	0.196
	70	2.211	-0.275	1.265	0.095	0.476	2.280	-0.087	1.720	0.186	0.442
Integrity	15	1.659	-0.017	6.069	0.588	0.218	1.622	0.129	3.511	0.648	0.204
	20	1.467	-0.095	3.944	0.544	0.257	1.838	0.042	3.315	0.574	0.241
	70	1.886	-0.251	1.293	0.023	0.514	2.233	-0.104	1.435	0.157	0.467

Table 3. Cluster Evaluation Results for HIV

Cluster Algorithm	#	Glasso					MB				
		Davies-Bouldin	Silhouette	Calinski-Harabasz	Baker-Hubert	Hubert-Levine	Davies-Bouldin	Silhouette	Calinski-Harabasz	Baker-Hubert	Hubert-Levine
Louvain	15	3.755	-0.065	1.039	0.016	0.547	1.554	0.306	10.891	0.700	0.167
	20	2.387	0.154	10.297	0.578	0.215	2.132	0.164	6.135	0.514	0.270
	70	3.657	0.003	3.146	0.250	0.414	2.637	0.012	2.145	0.202	0.434
Leiden	15	1.600	0.226	20.264	0.747	0.155	1.673	0.292	11.231	0.637	0.189
	20	2.350	0.162	10.774	0.606	0.203	2.150	0.168	6.123	0.481	0.280
	70	3.635	0.016	2.952	0.229	0.431	2.702	0.009	2.151	0.184	0.444
VAT	15	1.422	0.373	37.268	0.783	0.092	1.412	0.098	9.461	0.127	0.353
	20	2.764	-0.214	5.092	0.160	0.405	1.474	0.078	6.408	0.154	0.344
	70	2.952	-0.291	1.332	-0.105	0.576	2.168	-0.156	1.201	-0.100	0.577
Integrity	15	1.364	0.052	12.457	0.732	0.163	1.511	0.246	7.239	0.737	0.155
	20	1.871	-0.142	4.400	0.354	0.328	1.595	0.163	4.405	0.663	0.202
	70	2.713	-0.278	1.304	0.008	0.539	2.168	-0.156	1.201	-0.100	0.577

Leiden-20 graph for self esteem, the MB-VAT-20 graph for alcohol risk, and the Glasso-VAT-15 graph for HIV status. The clusterings are described below.

4.2 Cluster Properties

For the top three clusterings, the composition of each cluster for 40 variables, or attributes, is shown in Table 4. The color at the top of the column corresponds to the node color in the network's visualization. Each line of the table represents an attribute, and numbers shown are the percentages of cluster members displaying that attribute. Some variables, like *Income: 0 to 19999* and *In a Relationship* have similar percentages across most of the clusters. On the other hand, the percentage of cluster members who were previously diagnosed with syphilis (*STD: Syphilis*) varies greatly across clusters from 0 to 35%.

Self-Esteem Results for the high self esteem cohort are visualized in Fig 1. Here probands are divided into 5 clusters, displayed in violet, orange, green,

Table 4. Cluster composition for 40 variables. Numbers represent the percentage of cluster members exhibiting the described attribute. *Alc.* represents alcohol abuse risk.

	Alc.		**HIV+**			**Self Esteem**				
Variables / Cluster Number	0	1	0	1	2	0	1	2	3	4
Age: 18 to 30	53	100	17	3	5	36	31	28	15	50
Age: 31 to 45	43	0	58	71	82	65	62	46	86	50
Age: 46 to 99	6	0	24	26	14	0	8	28	0	0
Income: 0 to 19999	65	89	84	55	100	58	54	82	72	75
Income: 35000 to 75000	6	11	5	3	0	15	8	10	15	0
Living State: Relatives	8	22	5	3	14	15	24	10	0	0
Living State : Alone	18	22	33	58	5	22	24	10	29	25
Living State: Hotel	3	0	7	0	23	8	8	28	15	0
Living State: Roommates (Friends)	34	0	29	14	36	29	23	9	29	25
Education: Did Not Finish High School	57	56	50	60	77	37	48	65	87	50
Education: At Least Some College	36	46	43	34	18	59	47	29	15	50
In A Relationship	48	67	40	47	55	29	54	64	43	25
Talks To Friends: At Least Monthly	15	12	13	14	22	16	31	38	15	25
Talks To Friends: Yearly	38	11	41	37	27	29	54	37	43	25
Talks To Friends: Never	45	78	44	45	41	50	16	28	43	50
Talks To Family: At Least Monthly	29	24	43	46	27	23	16	19	29	25
Talks To Family: Yearly	26	23	12	29	14	43	31	37	29	0
Talks To Family: Never	29	56	18	11	32	8	31	19	29	75
Feels There Is No One To Turn To: Most Time	16	0	12	20	19	36	39	10	58	0
Feels There Is No One To Turn To: Sometimes	53	78	53	46	41	36	47	55	15	25
Feels There Is No One To Turn To: Never	21	22	25	29	23	0	0	10	15	50
Feels Left Out: Always	5	0	6	15	10	8	8	37	29	25
Feels Really Understood: At Least Most Time	56	0	49	40	49	87	78	83	58	50
Feels Really Understood: Never	14	0	21	23	19	0	0	10	0	0
Has Trusted Friend: At Least Most Time	54	0	41	41	47	87	62	83	86	25
Has Trusted Friend: Never	19	33	31	26	32	0	0	0	0	25
Has A Person To Lend Money: Always	8	0	16	5	5	29	8	28	0	0
Has A Person To Lend Money: Sometimes	24	56	22	15	23	22	16	10	15	50
Feels Part Of A Group: At Least Most Time	34	12	39	26	41	58	62	65	58	0
Feels Part Of A Group: Never	34	67	28	40	15	8	0	19	29	50
Feels Lack Of Companionship: Always	11	22	3	6	5	36	16	10	15	0
Feels Lack Of Companionship: Never	18	33	26	26	10	15	16	10	0	75
STD: Syphilis	13	0	39	12	5	0	24	19	15	0
STD: Gonorrhea	15	11	40	3	5	15	31	19	0	25
STD: Genital warts	15	11	36	3	5	22	16	28	0	25
STD: Genital Herpes	8	0	28	3	5	36	0	0	0	25
STD: HIV	17	0	100	100	100	50	31	28	29	25
STD: None	46	67	1	84	100	22	47	46	58	50
Unprotected Sex (last 2 months)	13	22	13	3	0	22	24	10	29	50
Unprotected Sex (last 12 months)	41	67	30	20	24	43	62	28	29	50
Size Of The Cluster (number of nodes)	119	9	113	35	22	14	13	11	7	4

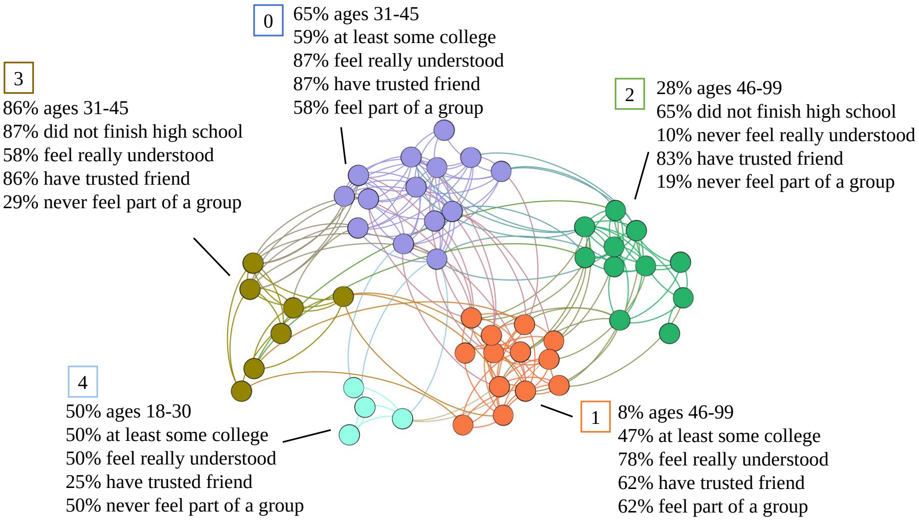

0
65% ages 31-45
59% at least some college
87% feel really understood
87% have trusted friend
58% feel part of a group
3
86% ages 31-45
87% did not finish high school
58% feel really understood
86% have trusted friend
29% never feel part of a group
2
28% ages 46-99
65% did not finish high school
10% never feel really understood
83% have trusted friend
19% never feel part of a group
4
50% ages 18-30
50% at least some college
50% feel really understood
25% have trusted friend
50% never feel part of a group
1
8% ages 46-99
47% at least some college
78% feel really understood
62% have trusted friend
62% feel part of a group

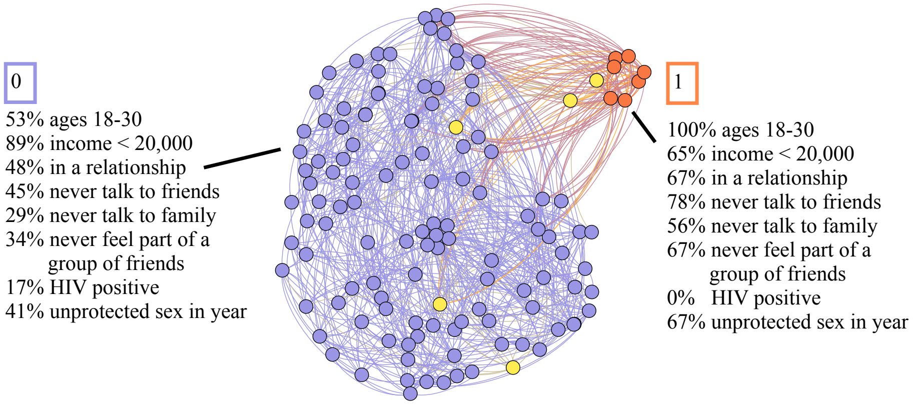
Fig. 1. Probands with high self-esteem are clustered.

brown, and arctic blue. Cluster 0 is 65% middle-aged, and is well-educated, with 59% of its members having at least some college. This cluster is distinguished by the prevalence of members who feel really understood and have a trusted friend (87%). More than half (58%) of its members feel part of a group of friends at least most of the time. Cluster 1 is somewhat older. 62% of members have a trusted friend and feel part of a group at least most of the time. 47% of its members have at least some college.

Cluster 2 is dominated by 83% of members who have a trusted friend; at the same time, 10% of its members never feel really understood, and 19% never feel part of a group of friends. This cluster is less educated, with 65% not finishing high school. Cluster 3 is distinguished by the middle-aged group, and a large percentage of members who have a trusted friend. More than half (58%) feel really understood, but 29% never feel of part of a group. Cluster 4 is the youngest group, half of whose members have at least some college and feel really understood.

Risk of Alcohol Abuse The results for risk of alcohol abuse are visualized in Fig 2 with two clusters in violet and orange. The clustering method is VAT, which produces an attack set, members of which are shown in yellow. Cluster 0 represents the majority of members, most of whom have a low income level. The group is older than cluster 1, with only 53% ages 18-30. 48% are in a relationship. Almost half of the members never talk to friends. Additionally, 17% of members are HIV positive, and 41% have had unprotected sex in the previous year.

Cluster 1 is a younger group, comprised entirely of 18 to 30 years old, 67% of whom are in a relationship. The cluster is more isolated than cluster 0 –



0

53% ages 18-30
89% income < 20,000
48% in a relationship
45% never talk to friends
29% never talk to family
34% never feel part of a group of friends
17% HIV positive
41% unprotected sex in year

1

100% ages 18-30
65% income < 20,000
67% in a relationship
78% never talk to friends
56% never talk to family
67% never feel part of a group of friends
0% HIV positive
67% unprotected sex in year

Fig. 2. Probands at risk for alcohol abuse are clustered.

78% never talk to friends and more than half never talk to family. 67% had unprotected sex in the last year, but none were HIV positive.

HIV+ Status The results for HIV positive status are visualized in Fig 3. The attack set nodes provided by VAT clustering (shown in yellow) along with information from the visualization allowed for enhanced analysis by identifying subgroups for clusters 1 and 2. Cluster 0 is the largest and youngest cluster, with 17% of members aged 18-30. Other clusters consist of members aged 31+. This cluster is primarily distinguished by the prevalence of STDs. For example, 39% have had syphilis and 40% have had gonorrhea. The prevalence of unprotected sex in this group is relatively high at 30% in the last year, and the group is sexually very active, with 13% having unprotected sex within the last 2 months.

Cluster 1 is divided into 4 sub-clusters. Cluster 1b, with many edges into cluster 0, is the only sub-cluster not free of STDs, with only 38% never reporting an STD diagnosis. Cluster 1d, which is attached to the graph by a single edge, truly consists of outsiders. None are in a relationship, and 100% live alone. They have very low rates of unprotected sex. Clusters 1a and 1c show no STDs, but have higher rates of unprotected sex at 38% and 29% within the last year.

Cluster 2 is 100% without STDs, and has relatively low rates of unprotected sex. The subclusters are distinguished primarily by living situation, with cluster 2a living in a hotel, and cluster 2b living with relatives, roommates, or friends.

5 Discussion and Conclusion

The clustering was able to find meaningful subgroups in the data, and it is particularly interesting that different variable combinations were important for the three targets. It is well-known that feature selection improves machine learning

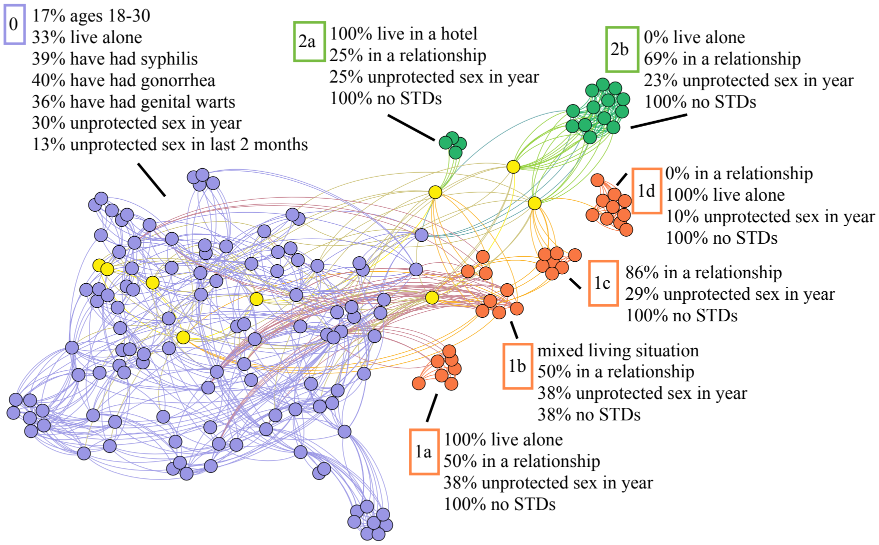


Fig. 3. Probands with HIV positive status are clustered.

performance, and this study confirms that result. Our best performing data was reduced from hundreds of variables to 15 or 20.

It is hard to pinpoint what factors lead to higher self-esteem; however results show that feeling really understood, having trusted friends, and belonging in a group of friends might lead to higher self-esteem. While these variables were relatively consistent, clusters were distinguished by age and education level.

For risk of alcohol abuse, the important variables were relationships and frequency of communication with family and friends. We identified a younger, more isolated group with a less developed social network and a higher propensity to engage in unprotected sex. From a public health perspective, efforts to prevent HIV targeting this group could also productively attempt to curb alcohol abuse. This is substantiated by many studies linking substance abuse and HIV [8].

For HIV+ status, one cluster had previous history of STDs, and two clusters did not. The connection between HIV and other STDs has been widely studied (e.g. see [5]), and this is another example where public health prevention efforts could be combined. The no-STD clusters were further distinguished by living situation and the frequency of unprotected sex.

From a network science perspective, graphs with the best internal evaluation scores were inferred with Glasso and MB. These methods were developed to work with microbial count data, which is similar to the attribute counts of the LMSM-CI dataset. There are previous examples of Glasso networks being used with medical data, such as to study PTSD [9] and frailty in older persons [12], but none that we could find with MB, suggesting that this is an area for further research. This result is also surprising in that Glasso and MB performed better than kNN, which is often used to infer graphs in data science applications. For clustering, it is said in [17] that NBR-Clust with VAT is useful for "initial exploratory clustering" and "where the number of desired clusters is low." Our study confirms that result, as the NBR-Clust with VAT clustering was able to find distinctive, useful clusters. Leiden also performed well. Network science based machine learning and analysis techniques were able to produce results of interest to workers in public health and other health-related areas.

References

1. Abbas, M., Matta, J., Le, T., Bensmail, H., Obafemi-Ajayi, T., Honavar, V., El-Manzalawy, Y.: Biomarker discovery in inflammatory bowel diseases using network-based feature selection. PloS one **14**(11), e0225,382 (2019)
2. Abirami, M., Vennila, B., Chilukalapalli, E.L., Kuriyedath, R.: A classification model to predict onset of smoking and drinking habits based on socio-economic and sociocultural factors. Journal of Ambient Intelligence and Humanized Computing **12**(3), 4171–4179 (2021)
3. Ahern, J., Karasek, D., Luedtke, A.R., Bruckner, T.A., van der Laan, M.J.: Racial/ethnic differences in the role of childhood adversities for mental disorders among a nationally representative sample of adolescents. Epidemiology **27**(5), 697–704 (2016)
4. Arbia, G., Bramante, R., Facchinetti, S., Zappa, D.: Modeling inter-country spatial financial interactions with graphical lasso: An application to sovereign co-risk evaluation. Regional Science and Urban Economics **70**, 72–79 (2018)
5. Barbee, L.A., Khosropour, C.M., Dombrowksi, J.C., Golden, M.R.: New hiv diagnosis independently associated with rectal gonorrhea and chlamydia in men who have sex with men. Sexually transmitted diseases **44**(7), 385 (2017)
6. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10,008 (2008)
7. Brugere, I., Gallagher, B., Berger-Wolf, T.Y.: Network structure inference, a survey: Motivations, methods, and applications. ACM Computing Surveys (CSUR) **51**(2), 1–39 (2018)
8. Bryant, K.J., Nelson, S., Braithwaite, R.S., Roach, D.: Integrating hiv/aids and alcohol research. Alcohol research & health **33**(3), 167 (2010)
9. Bryant, R.A., Creamer, M., O'Donnell, M., Forbes, D., McFarlane, A.C., Silove, D., Hadzi-Pavlovic, D.: Acute and chronic posttraumatic stress symptoms in the emergence of posttraumatic stress disorder: A network analysis. JAMA psychiatry **74**(2), 135–142 (2017)
10. CDC: Social determinants of health (2021). URL https://health.gov/healthypeople/objectives-and-data/social-determinants-health
11. Clifford, T., Bruce, J., Obafemi-Ajayi, T., Matta, J.: Comparative analysis of feature selection methods to identify biomarkers in a stroke-related dataset. In: 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1–8. IEEE (2019)
12. da Cunha Leme, D.E., da Costa Alves, E.V., Fattori, A.: Relationships between social, physical, and psychological factors in older persons: Frailty as an outcome in network analysis. Journal of the American Medical Directors Association **21**(9), 1309–1315 (2020)

13. Friedman, J., Alm, E.J.: Inferring correlation networks from genomic survey data. PLoS Comput Biol (2012)
14. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)
15. James, B.T., Luczak, B.B., Girgis, H.Z.: MeShClust: an intelligent tool for clustering DNA sequences. Nucleic Acids Research **46**(14), e83–e83 (2018). DOI 10.1093/nar/gky315. URL https://doi.org/10.1093/nar/gky315
16. Kino, S., Hsu, Y.T., Shiba, K., Chien, Y.S., Mita, C., Kawachi, I., Daoud, A.: A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. SSM-Population Health p. 100836 (2021)
17. Kramer, J., Boone, L., Clifford, T., Bruce, J., Matta, J.: Analysis of medical data using community detection on inferred networks. IEEE Journal of Biomedical and Health Informatics **24**(11), 3136–3143 (2020)
18. Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., Bonneau, R.A.: Sparse and compositionally robust inference of microbial ecological networks. PLoS computational biology **11**(5), e1004,226 (2015)
19. Marchette, D.J., Marchette, M.D.J., Matrix, S.: Package 'cccd'. https://cran.r-project.org/web/packages/cccd/index.html (2015)
20. Matta, J., Ercal, G., Sinha, K.: Comparing the speed and accuracy of approaches to betweenness centrality approximation. Computational Social Networks **6**(1), 1–30 (2019)
21. Matta, J., Obafemi-Ajayi, T., Borwey, J., Sinha, K., Wunsch, D., Ercal, G.: Node-based resilience measure clustering with applications to noisy and overlapping communities in complex networks. Applied Sciences **8**(8), 1307 (2018)
22. Matta, J., Zhao, J., Ercal, G., Obafemi-Ajayi, T.: Applications of node-based resilience graph theoretic framework to clustering autism spectrum disorders phenotypes. Applied network science **3**(1), 1–22 (2018)
23. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the lasso. The annals of statistics **34**(3), 1436–1462 (2006)
24. NIAAA: Drinking levels defined (2017). URL https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking
25. Parente, E., Zotta, T., Faust, K., De Filippis, F., Ercolini, D.: Structure of association networks in food bacterial communities. Food microbiology **73**, 49–60 (2018)
26. Phan, D.V., Yang, N.P., Kuo, C.Y., Chan, C.L.: Deep learning approaches for sleep disorder prediction in an asthma cohort. Journal of Asthma **58**(7), 903–911 (2021)
27. Ramirez-Valles, J.: The protective effects of community involvement for hiv risk behavior: a conceptual framework. Health Edu. Research **17**(4), 389–403 (2002)
28. Seligman, B., Tuljapurkar, S., Rehkopf, D.: Machine learning approaches to the social determinants of health in the health and retirement study. SSM-population health **4**, 95–99 (2018)
29. Traag, V., Waltman, L., van Eck, N.J.: From louvain to leiden: guaranteeing well-connected communities. arXiv preprint arXiv:1810.08473 (2018)
30. Udrescu, L., Sbârcea, L., Topîrceanu, A., Iovanovici, A., Kurunczi, L., Bogdan, P., Udrescu, M.: Clustering drug-drug interaction networks with energy model layouts: community analysis and drug repurposing. Scientific reports **6**(1), 1–10 (2016)
31. Walesiak, M., Dudek, A.: clustersim package. University of Wrocław, Wrocław http://keii. ue. wroc. pl/clusterSim (2010)
32. Yufang, T., Xueming, L., Yan, X., Shuchang, L.: Group lasso based collaborative representation for face recognition. In: 2014 4th IEEE International Conference on Network Infrastructure and Digital Content, pp. 79–83. IEEE (2014)