2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Graph Analytics for Healthcare Fraud Risk Estimation

L. Karl Branting, Flo Reeder, Jeffrey Gold, and Timothy Champney
The MITRE Corporation
7515 Colshire Dr
McLean, VA, 22102, USA
{lbranting, freeder, jgold, tchampney}@mitre.org

***Abstract*—This paper presents a novel approach to estimating healthcare fraud (HCF) risk that applies network algorithms to graphs derived from open source datasets. One group of algorithms calculates behavioral similarity to known fraudulent and non-fraudulent healthcare providers with respect to measurable healthcare activities, such as medical procedures and drug prescriptions. Another set of algorithms estimates propagation of risk from fraudulent healthcare providers through geospatial collocation, i.e., shared practice locations or other addresses. The algorithms were evaluated with respect to their ability to predict a provider's presence on the Office of the Inspector General's list of providers excluded from participation in Medicare and other Federal healthcare programs (*exclusion*). In an empirical evaluation, a combination of 11 features achieved an f-score of 0.919 and a ROC area of 0.960 in exclusion prediction. An ablation analysis showed that most of this predictive accuracy was the result of features that measure risk propagation through geospatial collocation.**

I. Introduction

Healthcare fraud (HCF) is a multibillion-dollar drain on healthcare spending [11], consuming an estimated $98 billion of annual Medicare and Medicaid spending in the United States.[1] The magnitude of HCF is very large in proportion to the resources available for investigation and prosecution of these fraudulent activities, making prioritization of investigative leads essential. Automated estimation of HCF risk has the potential to maximize the use of limited investigative resources by identifying the individuals and institutions at highest risk of fraud.

Graph analysis is a promising framework for HCF risk assessment for several reasons. First, HCF often involves multiple entities; by making relationships among such entities explicit, graph representations facilitate algorithms for detecting coordinated activity and the spread of social influence. In addition, graph analytics have a proven track record in law enforcement and intelligence analysis applications, suggesting that they might be equally useful in the HCF domain.

This paper describes an initial exploration of graph analytics for HCF risk assessment using open source healthcare data. We describe the modeling and ingestion of open source datasets and propose novel algorithms for predicting an observable consequence of HCF: presence on the list of providers excluded from participation in Medicare (and other Federal healthcare programs) published monthly by the Office of the Inspector General (OIG). We refer to presence on this list as *exclusion* and the likelihood that a given provider will be on the OIG list as the provider's *exclusion risk*.

II. Datasets

Healthcare providers excluded from Medicare eligibility are set forth in a "List of Excluded Individuals/Entities" (LEIE) published by the OIG.[2] Unfortunately, the LEIE comprises only a small subset of individuals identified by insurance providers or law-enforcement officials as having committed HCF. For example, providers who are accused of overcharging insurers or Medicare often relinquish the overpayments without any public acknowledgement or notice. However, the LEIE is among the few open source data sets of HCF information. We therefore treat inclusion on the LEIE as the predicted variable whose probability we strive to estimate as an observable indication of HCF.

We draw our predictive variables from three datasets. The first is the "Medicare Provider Utilization and Payment Data: Physician and Other Supplier" (PUF) data for 2012, 2013, and 2014 published by the Centers for Medicare & Medicaid Services (CMS).[3] This summarizes each provider's annual charges to Medicare under each treatment category denoted by a Healthcare Common Procedure Coding System (HCPCS) code. We hypothesized that excluded providers may have distinctive billing patterns that distinguish them from non-excluded providers.

The second dataset of predictive data consists of the "Medicare Provider Utilization and Payment Data: Part D Prescriber" (Part-D) data for 2013 published by CMS.[4] This consists of information about the prescription drugs that individual physicians and other healthcare providers prescribe in the Medicare Part D Prescription Drug Program. We hypothesize that excluded and non-excluded providers may differ in their drug-prescribing patterns.

Finally, we used National Provider Identifiers (NPIs) to obtain unique identifiers for providers. NPIs are set forth in the National Plan and Provider Enumeration System (NPPES) data set from 2015.[5] Unfortunately, the LEIE has NPIs for only 5% of the excluded providers, in part because a large proportion of providers on the list were excluded prior to NPI requirements. We therefore found it necessary to implement an identity-matching procedure in Lucene[6] to match excluded providers in the LEIE with providers in the NPPES.

IEEE/ACM ASONAM 2016, August 18-21, 2016 San Francisco, CA, USA
978-1-5090-2846-7/16/$31.00

The identity-matching algorithm compared names in the LEIE of both organizations and individual providers after we preprocessed the NPPES and LEIE datasets for variations in name conventions. For instance, LEIE could list an organization as "St Joe" when NPPES would have "St. Joe." Organizations, in particular, showed varied name conventions. To account for potential spelling mistakes, the matching procedure looks not just for exact matches, but also for matches based on phonetic similarity or edit distance. The algorithm used additional features for identity matching, including requiring that a matching set of providers have addresses in the same state. We evaluated the accuracy of identity matching on a set of providers whose NPIs were specified in the LEIE dataset. In a leave-one-out evaluation, the algorithm found 82% of the providers in the test set.

Even with this level of performance, only 10-15% of LEIE providers with missing NPIs could be matched with sufficient certainty, so the number of known true positives, that is, providers known to be excluded (just 12,153 providers matched providers at least one drug prescription, treatment code, or location), is small relative to the total number of providers (over 4.7 million).

III. Graph Design and Implementation

For the experiments described below, we implemented our graph in Neo4j,[7] a popular open source graph database, using the 2.3.0 baseline. However, nothing in the experiment is dependent on this particular choice of graph database; we also implemented several of the graph algorithms in Spark GraphX,[8] which may be a more suitable platform for extremely large graphs because, unlike Neo4j, GraphX can operate on distributed architectures.

Creation of a graph database is a modeling activity in that there is no unique graph representation of any particular set of data; the best representation is one that best facilitates the particular algorithms of interest. In our case, we wished to evaluate the feasibility of estimating HCF risk by comparing providers based on billing and drug-prescription behavior and by estimating risk propagation through location links. We therefore represented each provider, each prescription drug, and each treatment code (HCPCS) as a separate node. Providers were linked to each drug they prescribed by a PRESCRIBED edge whose attributes represented the number of patients prescribed the drug by the provider, the total cost of the prescriptions, and other Part-D data. Providers were linked to each HCPCS by a CHARGE_OF link whose attributes include average Medicare payment, the number of unique beneficiaries, and other treatment information that appears as columns in the PUF dataset .

To facilitate graph-based geospatial reasoning, locations were represented as nodes, and each provider was connected by LOCATED_AT edges to each address associated with that provider. In general, there could be multiple such addresses corresponding to multiple practice locations or institutional affiliations. Addresses in these files were run through a geospatial tagging (geotag) algorithm which yields latitude and longitude for each provider address. For some addresses only city and state were used since street addresses were not available. LOCATION nodes are uniquely identified by a combination of latitude and longitude. Other nodes included exclusions (there are 17 distinct exclusion codes, each corresponding to a possible reason for being excluded) and generic drugs (distinct from, but sometimes equivalent to, non-generic drugs).

Figure 1 shows a small example subgraph illustrating the key nodes and relationships. This graph contains roughly 173 million edges and 5.05 million nodes.

IV. Graph Analytics

We distinguish three basic categories of graph analytics:

1) Similarity functions between pairs of entities based on structural similarity
2) Attribute estimation based on network propagation
3) Complex structure detection and analysis

Structural similarity analytics include alias detection [7] and nearest-neighbor classification or regression. In this work, we estimate behavioral-similarity from structural similarity. The second family of graph analytics, those based on network propagation, can be used to predict cascades or epidemics or to estimate centrality, status, or contagion [4], [2]. In this work, we estimate the exclusion risk of each individual provider based on network flow to that provider from known excluded providers through location edges. The third category of graph analytics is not addressed in this work because we lack *a priori* models of fraudulent enterprises, organizations, or other complex structures.

A. Behavior-Vector Similarity

As described above, we hypothesized that the billing or drug-prescription behavior of providers might predict exclusion risk. Our framework for behavior-based risk estimation is shown in Figure 2. We compare each unknown provider to each member of positive and negative reference sets. The similarity of each pair of providers, in turn, is determined by comparing the pivot nodes adjacent to each, where pivots are either treatment types (HCPCS) or drugs.

Specifically, for each provider (unknown or reference-set member) we define a *behavior-vector* consisting of a vector of edge weights to each adjacent pivot. The value of each term in the vector corresponding to a given pivot is either 0.0, if the provider node has no edge to that pivot, or a value calculated in a manner specific to the type of pivot. In our experiments, the weight of an edge to a drug-type node is the "TotalDrugCost," meaning the total amount billed by this provider for this particular drug (this value is a column in the Part-D dataset). The weight of an edge to a treatment type (HCPCS) is calculated as the product of the "AvgMedicarePayment" and the "BeneUniqCount," which are columns in the PUF dataset that represent the average Medicare payment and the number of unique beneficiaries, respectively. The similarity between any pair of providers with respect to a given form of behavior (such as drug prescription or billing) can be estimated by

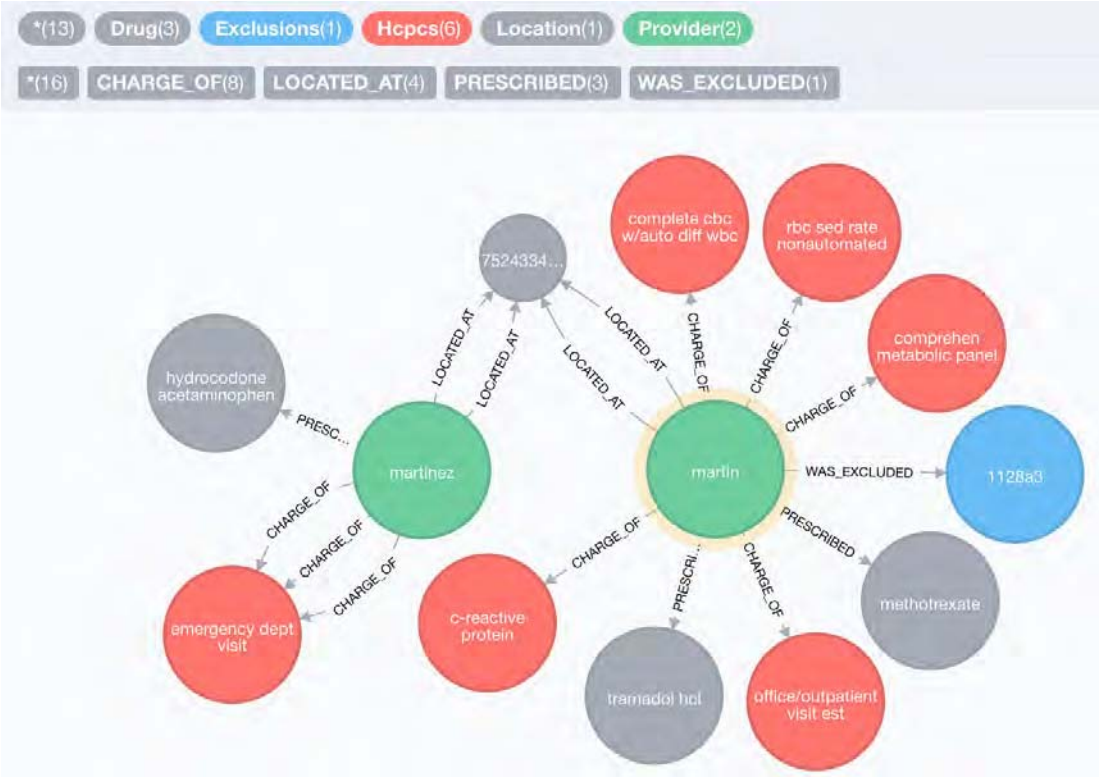

Fig. 1. Neo4j web server view of a subgraph showing a provider, labeled "martinez," who shares a location with an excluded provider,"martin," together with some of these providers' prescriptions and HCPCS (treatment codes).

applying a standard vector-similarity metric to the providers' behavior vectors.

There are numerous vector-similarity metrics, including Jaccard, Dice, and L norms, but we obtained the best results using cosine similarity, which produces meaningful results even when the vectors being compared differ significantly in magnitude. In making the cosine calculation, we experimented with scaling each term associated with a given pivot by the log-inverse vertex degree of that pivot. This formalizes the intuition that the fact that two people both watch the same Super Bowl game (a very common connection) says much less about their similarity than if they watch the same TED Talk (a much rarer connection) or the fact that two people both live in Los Angeles says less about their similarity than if they both live in Taos or Nome. We observed that inverse vertex-degree weighting was generally associated with better performance.

To convert the inverse-degree-weighted cosine between a provider and members of positive and negative reference sets into features usable for risk estimation, we took the mean of the k closest members of each set, producing three features:

1) **negative-similarity**, mean similarity to the closest k members of the negative reference set
2) **positive-similarity**, mean similarity to the closest k members of the positive reference set
3) **neg-sim-ratio**, the ratio of features 1 and 2

Repeating this calculation for both types of pivots produces 6 real-valued features: HCPCS-negative-similarity, HCPCS-positive-similarity, HCPCS-ratio, drug-negative-similarity, drug-positive-similarity, and drug-ratio.

B. Risk Propagation

We estimated propagation of risk through collocations, i.e., shared addresses. collocation was based on edges connecting providers to practice, mailing, or business addresses. We selected collocation as a potential feature based on input from fraud investigators who described criminal networks where fraud was endemic to geographic areas, such as the pill mills in Florida[9].

Four approaches were explored to measure risk propagation through location edges.

1) **bad-colocator-count**, the number of collocated excluded providers
2) **bad-2hop-colocator-count**, the number of paths to excluded providers two location hops away
3) **bad-colocator-ratio**, the number of collocated excluded providers, scaled by the vertex degree of each location (so that a high-degree location contributes less to the score than a low-degree location)
4) **bad-2hop-colocator-ratio**, the number of paths to excluded providers two location hops away, scaled by the vertex degree of each location

For example, in Figure 3 there are 4 paths from provider1 to positive instances, so bad-colocator-count = 4, while the bad-

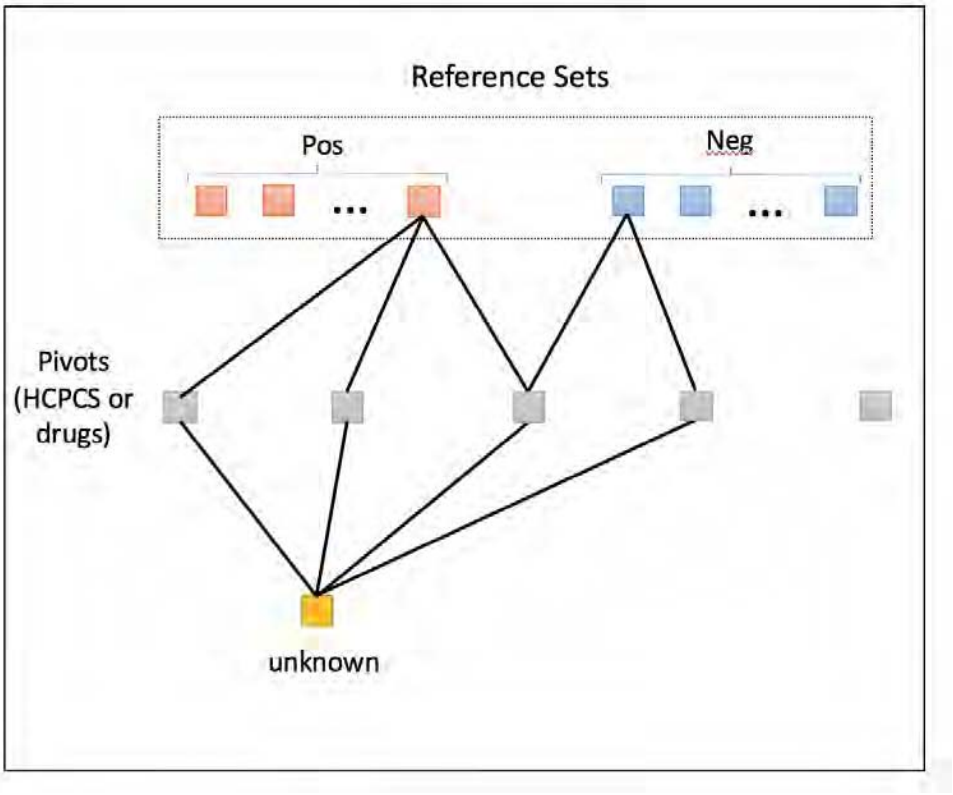

Fig. 2. A framework for behavior-based risk estimation in which a provider is compared to known positive and negative instances with respect to weighted connections to *pivots*, which are nodes that may be indicative of behavior, such as treatment types (HCPCS) or drug prescriptions, and that are common to multiple providers.

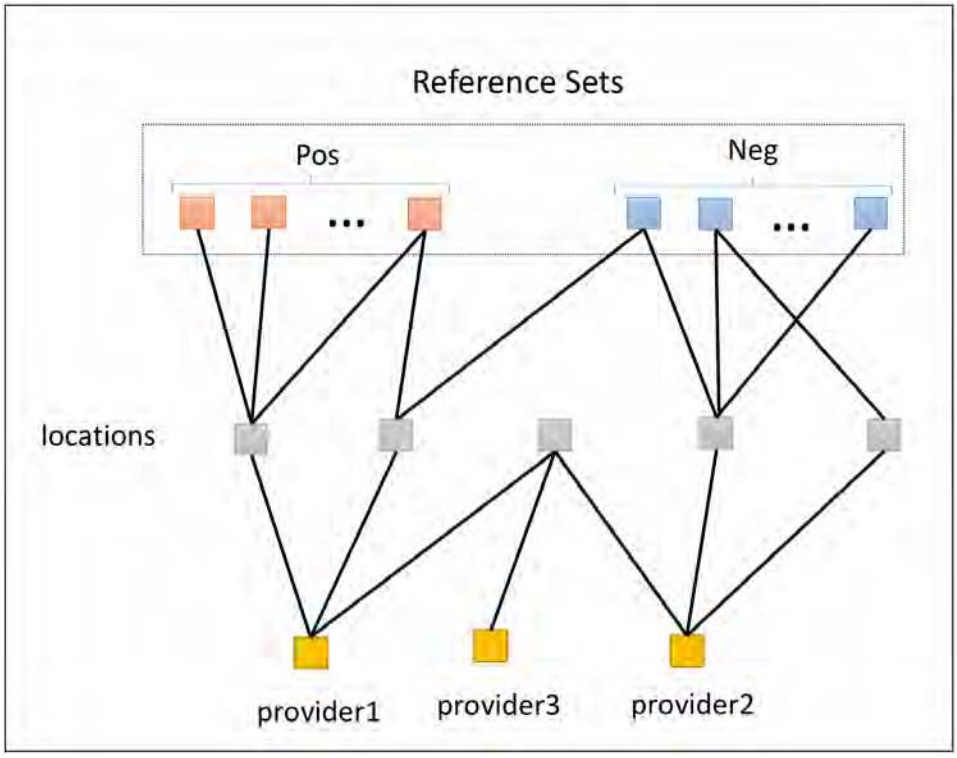

Fig. 3. Risk propagation is estimated by counting the number of distinct paths from a given provider to excluded providers (labeled ”Pos” for ”positive instance” in the figure).

colocator-count of provider2 = 0. In provider1's bad-colocator-ratio, the number of paths through each location node is scaled by the vertex degree of the location node, i.e., $3/3+1/2 = 1.5$. Provider3's bad-2hop-colocator-count = 4 + 0 = 4 and bad-2hop-colocator-ratio = $1.5 + 0 = 1.5$.

C. Provider Attributes

Although the focus of this work was on graph analytics, we experimented with one provider attribute:

1) **new-patients-per-bene**, the proportion of services provided to new patients. The intuition is that a provider who renders a large number of services for each new patient may be rendering more services than medically necessary.

V. Experimental Design

We performed an experiment on a set of 12,000 excluded providers, each of which had at least one drug prescription, treatment, or address (i.e., a LOCATED_AT, CHARGE_OF, or PRESCRIBED link). Providers lacking these links would be disconnected from the graph and therefore not amenable to graph analysis. As discussed above, deriving exclusion-risk estimates based on drug or HCPCS behavior-vectors requires reference sets of known excluded and non-excluded providers. Accordingly, the 12,000 excluded providers and 12,000 randomly selected non-excluded providers were split into a reference set containing 6,000 reference excluded providers and 6,000 reference non-excluded providers, and a cross-validation set containing 6,000 test excluded providers and 6,000 test non-excluded providers.

Following the procedures described in the previous section, 11 features were derived for each member of the cross-validation set. In doing so we excluded pivots with vertex degree greater than 100,000 in calculating behavior-vector similarity, (on the assumption that such high-degree pivots have little discriminative ability and slow computation); 6429 HCPCS nodes and 2669 drug-type nodes were below this threshold. For the nearest-neighbor calculations, $k = 3$.

The 11 graph-based features consisted of 3 HCPCS structural-similarity features, 3 drug-type behavior-vector features, 1 feature based on provider attributes, and 4 risk-propagation features. These features were used as input to supervised concept-learning algorithms in the WEKA[10] framework. The results below were calculated applying 10-fold cross validation using WEKA's j48 decision tree algorithm. Similar results were obtained from other inductive algorithms, but j48 is quite fast, and decision tree models are particularly easy to understand.

VI. Results

In 10-fold cross validation on the full 12,000 member 11-feature dataset, the mean f-measure was 0.919 and the mean ROC area was 0.960. This result indicates that graph features distinguish excluded providers from non-excluded providers with an impressive degree of accuracy.

TABLE I
CONTRIBUTIONS OF VARIOUS GRAPH-ANALYTIC FEATURE SUBSETS TO OVERALL 10-FOLD CROSS-VALIDATION ACCURACY.

feature set	f-measure	ROC area
All	0.919	0.960
Risk propagation features	0.901	0.815
HCPCS behavior-similarity features	0.715	0.718
Provider attributes	0.697	0.697
Drug behavior-similarity features	0.657	0.666

TABLE II
MEAN CLASSIFICATION ACCURACY IN 10-FOLD CROSS-VALIDATION OF FOUR TYPES OF RISK-PROPAGATION FEATURES

feature set	f-measure	ROC area
coloc-1-hop	0.896	0.892
coloc-2-hop	0.571	0.619
coloc-scaled-1-hop	0.896	0.892
coloc-scaled-2-hop	0.479	0.569

To better understand the contributions of the various graph features and provider attributes, we performed an ablation study that separated the contribution to classification accuracy of each of 4 distinct subsets of the remaining features, as set forth in TableI and summarized in Figure 4:

These results indicate that by far the most predictive of the remaining features are those based on risk propagation through location edges. The union of all the features is nevertheless more predictive than any one feature type by itself.

Table II separates the performance of the four individual risk-propagation features, indicating that almost all of the contribution is from 1-hop collocation. Scaling the count of bad collocators by location vertex degree appears to have little effect on accuracy.

Another view of the relative contribution of the 11 graph-analytic-derived features is set forth in Table III, which shows the mutual information between each feature and the category "exclusion." One-hop collocation has the most information, but all the features are informative.

It is striking that risk propagation appeared to contribute so strongly to risk prediction, given that network propagation has been shown to be highly predictive of individual attributes in other contexts. We surmise that location is a strong proxy for

TABLE III
MUTUAL INFORMATION BETWEEN GRAPH-ANALYTIC FEATURES AND MEDICARE EXCLUSION

feature	mutual information
coloc-1-hop	0.5989
coloc-scaled-1-hop	0.5793
HCPCS-negative-sim	0.1667
new-patient-per-bene	0.1483
HCPCS-ratio	0.1265
HCPCS-positive-similarity	0.1244
drugs-negative-similarity	0.1034
coloc-scaled-2-hop	0.0989
coloc-2-hop	0.0956
drugs-ratio	0.0956
drugs-positive-similarity	0.0937

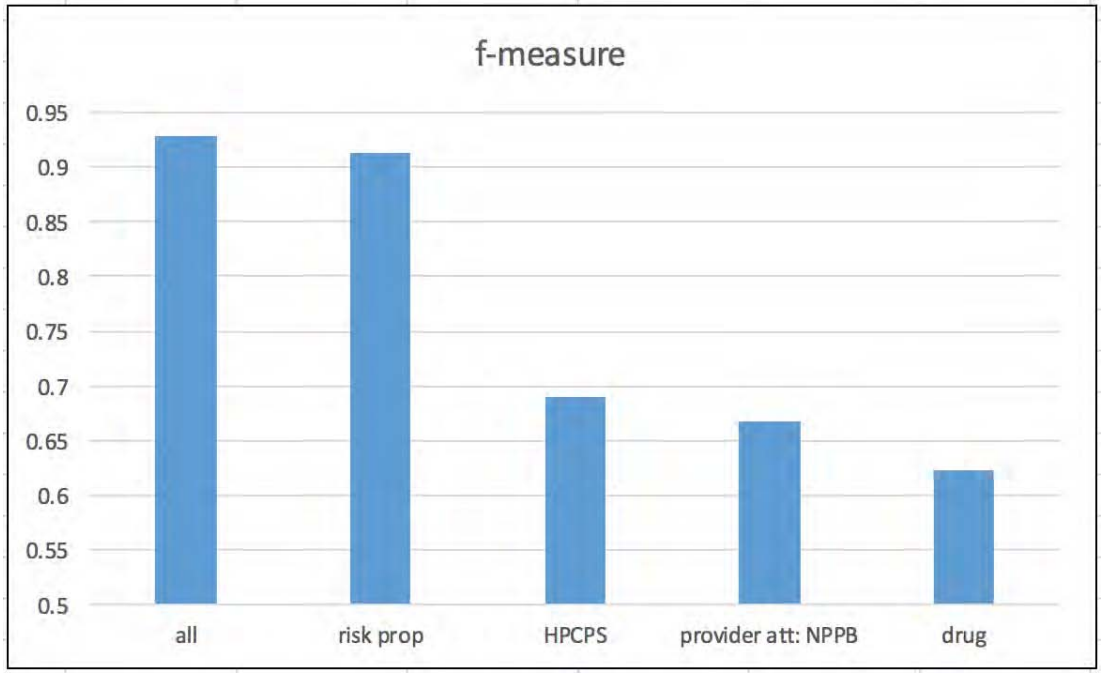

Fig. 4. The relative contributions of risk propagation, HPCPS, provider, and drug features to overall 10-fold cross-validation accuracy.

the kind of social connections that facilitate the propagation of social influence [5] .

To provide a baseline for comparison, we used a BayesNet classifier in WEKA. It was run with 10-fold cross-validation on a set of randomly selected excluded and non-excluded providers who had at least one HCPCS or PartD (prescription) charge. The data set size was 26455 providers. Because the provider pool is small, we could not use counts for individual HCPCS codes or drug types, so we aggregated individual HCPCS charges on the number of HCPCS instances and total service count and PartD charges on the number of charges and total prescription count for all years available. In addition, both business address latitude/longitude and practice address latitude/longitude served as a geographic indicators. The classifier predicted category correctly only 87.8% of the time with an ROC of 0.684. While the regression classifier was more successful, predicting risk 91% of the time the ROC was only 0.679.

VII. Related Work

There is a rich and rapidly expanding research literature on social network analysis.[11] Network analysis appears in crime detection such as [10] which examined social proximity of gang members and how this influenced the probability of being a gunshot victim. [6] describe several social network analysis efforts including using YouTube social graphs to detect terror networks.

Previous work in graph analytics for healthcare fraud detection has typically emphasized anomaly detection, e.g., [9], [1], [3], rather than supervised concept learning. The paucity of open source ground truth healthcare fraud datasets has made large-scale evaluations of HCF analytics, such as is set forth in this paper, rare.

The approach of measuring similarity between nodes by structural similarity, applied in this work to detect behavioral similarity with respect to treatment and drug-prescription patterns, has been used primarily for alias detection [7] and has generally not used cosine as a similarity metric. One recent exception [8] looked only at prescription abuse. A vast amount of research addresses propagation of influence through social networks, e.g., [5]. However, to the best of our knowledge, our work is the first to apply this approach for HCF risk assessment.

VIII. Summary and Future Work

This paper has described how the problem of predicting HCF risk can be formulated in terms of network algorithms operating on a graph derived from Medicare payment, location, drug-prescription, and exclusion data. An empirical evaluation demonstrated that a combination of 11 features can achieve an f-score of 0.919 and a ROC area of 0.960 in 10-fold cross-validation and that the most predictive features are based on collocation-based risk propagation.

We anticipate that a richer graph representation and more extensive target set would significantly improve the predictive accuracy of this approach. In particular, we used HCPCS and drug prescriptions as pivots in assessing behavioral similarity, but many other types of behavioral similarity might be equally or more informative. As we increase the richness of our graph, we expect to evaluate these alternative sources of behavior-similarity information and explore the extent to which combining these sources leads to improved predictive accuracy.

We plan to apply this work to datasets containing much more extensive true positives, that is, providers known to have committed healthcare fraud, and to richer graphs containing many additional types of information relevant to HCF prediction. In addition, improving the identity-matching algorithms to include other match types or relaxing match restrictions could provide a larger target set.

ACKNOWLEDGMENT

This work was funded under contract number CECOM W15P7T-09-C-F600. The MITRE Corporation is a private, not-for-profit corporation that operates Federally Funded Research and Development Centers in the public interest.

NOTES

[1] http://www.economist.com/news/united-states/21603078-why-thieves-love-americas-health-care-system-272-billion-swindle

[2] http://oig.hhs.gov/exclusions/exclusions_list.asp

[3] https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/medicare-provider-charge-data/physician-and-other-supplier.html

[4] https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber.html

[5] http://www.resdac.org/resconnect/articles/121

[6] https://lucene.apache.org/

[7] neo4.com

[8] http://spark.apache.org/graphx/

[9] https://www.washingtonpost.com/news/to-your-health/wp/2015/12/21/florida-pill-mill-crackdown-also-may-have-reduced-heroin-deaths-researchers-find/

[10] http://www.cs.waikato.ac.nz/ml/weka/

[11] See, for example, the journal of Social Network Analysis and Mining and the IEEE/ACM International Conferences on Advances in Social Networks Analysis and Mining

REFERENCES

[1] Akoglu, L., Faloutsos, C.: Anomaly, event, and fraud detection in large network datasets. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. pp. 773–774. WSDM '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2433396.2433496

[2] Aven, B.L.: The paradox of corrupt networks: An analysis of organizational crime at enron. Organization Science 26(4), 980–996 (2015)

[3] Chandola, V., Sukumar, S.R., Schryver, J.C.: Knowledge discovery from massive healthcare claims data. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1312–1320. KDD '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2487575.2488205

[4] Choobdar, S., Ribeiro, P., Parthasarathy, S., Silva, F.: Dynamic inference of social roles in information cascades. Data Mining and Knowledge Discovery 29(5), 1152–1177 (2015)

[5] Christakis, N.A., Fowler, J.H.: Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives – How Your Friends' Friends' Friends Affect Everything You Feel, Think, and Do. Back Bay Books (2009)

[6] Edwards, M., Rashid, A., Rayson, P.: A systematic survey of online data mining technology intended for law enforcement. ACM Computing Surveys 48(1) (September 2015)

[7] Hsiung, P., Moore, A., Neill, D., Schneider, J.: Alias detection in link data sets. In: Proceedings of the International Conference on Intelligence Analysis (May 2005)

[8] Liu, E., Wilson, A., Guerra-Gomez, J., Honda, T., Sricharan, K., Gilpin, L., Davies, D.: Graph analysis for detecting fraud, waste, and abuse in health-care data. AI Magazine 37(2) (2016)

[9] Liu, J., Bier, E., Wilson, A., Honda, T., Sricharan, K., Gilpin, L., Guerra-Gomez, J., Davies, D.: Graph analysis for detecting fraud, waste, and abuse in healthcare data. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA. pp. 3912–3919 (2015)

[10] Papachristos, A., Braga, A., Piza, E., Grossman, L.: The company your keep? the spillover effects of gang membership on individual gunshot victimization in a co-offending network. American Society of Criminology 53(4), 624649 (November 2015)

[11] Sparrow, M.: License to steal: how fraud bleeds America's health care system. Westview Press (2000)