Healthcare Billing Fraud Detection and Analysis Using Multiple Methods

Yue Li

Statistics
Duke University

04/05/2024

Background & Research Questions

Healthcare fraud is more prevalent among medical providers and usually results in higher health care costs, insurance premiums, and taxes for the general population. Medical Providers try to maximize reimbursement received from Medicare which they are not entitled to via illegitimate activities such as submitting false claims.

Common billing frauds :

- Duplicate billing
- Phantom billing
- Upcoding
- Unbundling

Research questions :

- What are some common patterns of billing frauds present ?
- What are some key characteristics of key fraud providers ?
- What are some key characteristics of patients susceptible to fraud ?
- What data-driven business recommendations can we make for fraud prevention ?

Dataset (from CMS.gov)

Inpatient/Outpatient data : claim-level data for patients who stayed at/didn't stay at the hospital for medical service.
Beneficiary data : information about patient who submitted claims.
Provider data : information about whether the provider is fraudulent.

Data	Column	Description
Outpatient & Inpatient	BeneID	Unique id of each patient
	ClaimID	Unique id of the claim submitted by the provider
	ClaimStartDt	Date when the claim started
	ClaimEndDt	Date when the claim ended
	Provider	Unique id of the provider
	InscClaimAmtReimbursed	Amount reimbursed for that particular claim
	AttendingPhysician	Id of the Physician who attended the patient
	OperatingPhysician	Id of the Physician who operated on the patient
	OtherPhysician	Id of the Physician other than AttendingPhysician and OperatingPhysician who treated the patient
	ClmDiagnosisCode 1-10	Codes of the diagnosis performed by the provider on the patient for that claim
	ClmProcedureCode 1-6	Codes of the procedures of the patient for treatment for that claim
	DeductibleAmtPaid	Amount by the patient before insurance plan (=Total claim amount — Reimbursed amount)
	ClmAdmitDiagnosisCode	Numerical Diagnosis Code on the claim indicating patient's initial diagnosis at admission
	DiagnosisGroupCode	Numerical Diagnosis Code assigned by the primary doctor according to diagnosis results
Inpatient	AdmissionDt	Date on which the patient was admitted into the hospital
	DischargeDt	Date on which the patient was discharged from the hospital
Beneficiary	BeneID	Unique id of the patient
	DOB/DOD	Date of Birth/Death of the patient
	Gender, Race, State, County	Demographics of the patient
	RenalDiseaseIndicator	Indicates if the patient has existing kidney disease
	IPAnnualReimbursementAmt	Maximum reimbursement amount for hospitalization annually
	IPAnnualDeductibleAmt	Premium paid by the patient for hospitalization annually
	OPAnnualReimbursementAmt	Maximum reimbursement amount for outpatient visits annually
	OPAnnualDeductibleAmt	Premium paid by the patient for outpatient visits annually
Provider	Provider	Unique id of the provider
	PotentialFraud	**Target**: whether the provider is fraudulent

After merging the data by providers, the whole dataset contains 558,211 claims and 5,410 providers.

Key Findings from EDA

Provider-related :

- There are considerably more fraudulent providers offering both outpatient and inpatient services than those offer single service.
- On average, the total claim amount for fraudulent providers is $2,500 more than non-fraudulent ones, and the daily total charge for fraudulent providers is $470 more than non-fraudulent ones.
- On average, fraudulent providers use 38 more group diagnosis codes and 65 more claim admit diagnosis codes than non-fraudulent ones.
- Outpatient providers duplicate 247 times own medical records than inpatient ones, and duplicate 7 times medical records from other providers than inpatient ones. 23% of fraudulent providers reuse patient IDs, and 29% of fraudulent providers reuse doctor IDs.
- Most of providers are local at state level. As the given number of states of a provider operates in increase, # non-fraudulent providers/# fraudulent providers decreases.

Patient-related :

- The highest number of claims were filed for patients with 4-6 chronic conditions.
- Patients aged 65-82 file the most claims. Patient aged 65-85 with diabetes and ischemic heart have the most claims, and most likely susceptible to fraud.

Data Preprocessing & Feature Engineering

Missing values and outliers :

- Impute Physician-related and Claim Procedure Code columns with `None`. Impute null values in numerical columns with 0.
- Keep outliers in reimbursement amounts.

Feature engineering :

Category	Column	Description
Financial	DeductibleAmtPaid	Average duration of patients stay in hospital by provider
	InscClaimAmtReimbursed	Average insurance claim amount reimbursed for patient per provider
	InscCoveredPercent	Average insurance coverage per provider
	TotalClaimAmount	InscClaimAmtReimbursed+DeductibleAmtPaid
	DailyTotalCharge	TotalClaimAmount/DaysAdmitted
Provider	DaysAdmitted	DischargeDt-AdmissionDt
	ServiceType	Type of service provided by provider (Inpatient/Outpatient/Both)
	numState	Number of states the provider operates in
	numCounties	Number of counties the provider operates in
Patient	NumOfPatients	Number of patients per provider
	NumChronicCond	Average number of chronic conditions per provider
	NumOfDuplicatedBeneID	Count of duplicated patient id per provider
	AvgAgeWhenServed	Average patient age per provider
Physician	NumOfDoctors	Number of doctors per provider
	NumOfDuplicatedAttendingPhysician	Count of duplicated attending physician per provider
	NumDistincOpPhy	Count of distinct operating physicians per provider
	NumDistincOtherPhy	Count of distinct other physicians per provider
Claim	NumOfClms	Number of claims filed per provider
	WeeklyClaims	Number of claims filed weekly per provider
	MonthlyClaims	Number of claims filed monthly per provider
	NumOfDuplicatedClaims	Count of duplicated claims filed per provider
	NumDistincClmProCode1-5	Count of distinct procedure diagnosis code 1-5 per provider
	NumDistincClmDiagCode1-10	Count of distinct claim admit diagnosis code 1-10 per provider
Code	numDiffDiagnosisCode	Number of claim admit diagnosis code per provider
	numDiffGroupDiagCode	Number of group diagnosis code per provider

Remove features : all claim-related and diagnosis-related codes (a lot of NAs), features from which other features are created, and perform feature selection through random forest and LASSO.

43 features remain after removal.

Predictive Modeling : Classification with Umbalanced Target

Models : Logistic Regression, AdaBoost, Random Forest, XGBoost, LightGBM.
Expectations : good model interpretations, low FNR, low FPR if possible.
Steps : Train-test split $\rightarrow$ Upsample the minority class via SMOTE $\rightarrow$ Standardize the features $\rightarrow$ GridSearch for the best parameters $\rightarrow$ Train the models $\rightarrow$ Model evaluation $\rightarrow$ Interpretation
Evaluation :

Model	Precision	Recall	F1 Score	AUC Score	FPR	FNR
Logistic	0.825	0.478	0.606	0.689	25.79%	5.39%
AdaBoost	0.692	0.664	0.678	0.684	3.25%	38.50%
RandomForest	0.805	0.718	0.759	0.772	13.59%	9.22%
XGBoost	0.895	0.821	0.856	0.862	8.95%	4.83%
LightGBM	0.906	0.828	0.866	0.871	8.58%	4.28%

Conclusion : LightGBM outperforms other models. However, logistic regression has a surprisingly low FNR.

Interpretation : Feature Importance

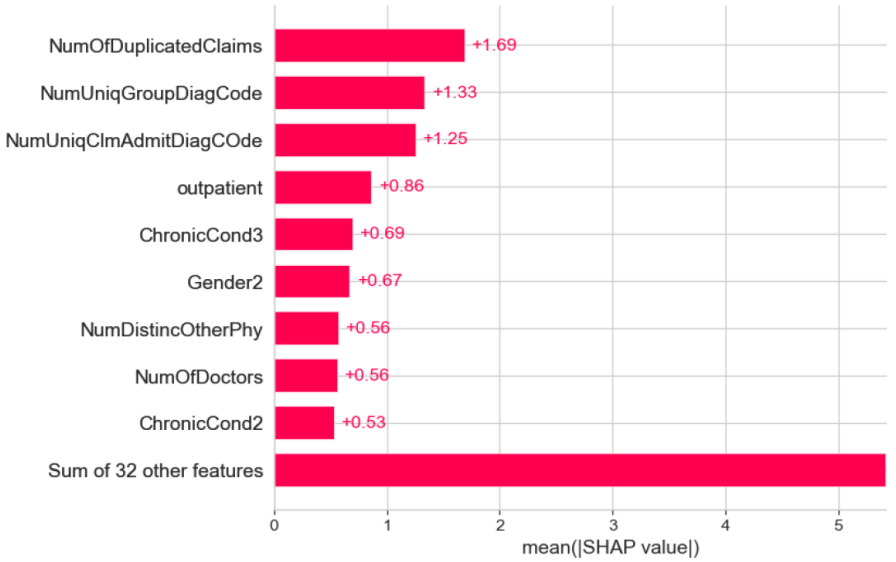


Figure 16: Logistic Regresion SHAP values

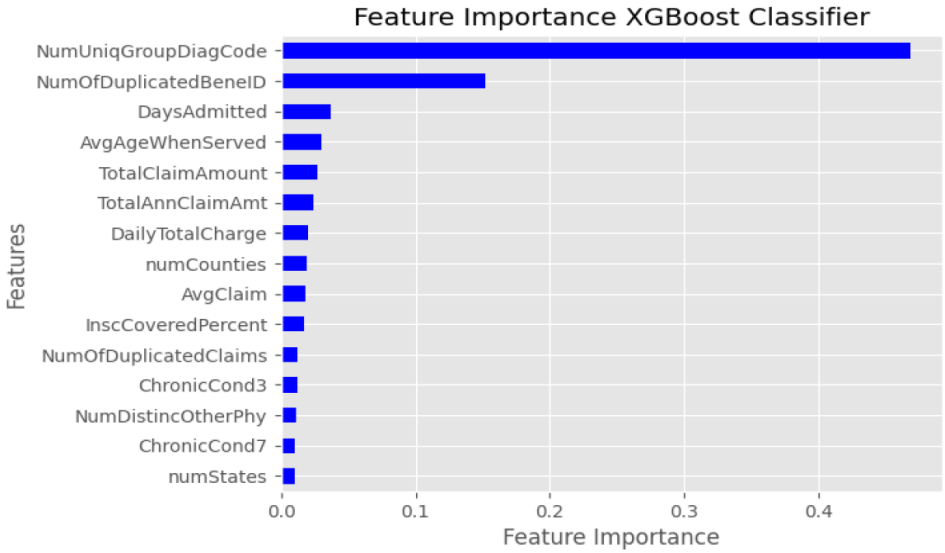


Figure 17: XGBoost Feature Importance

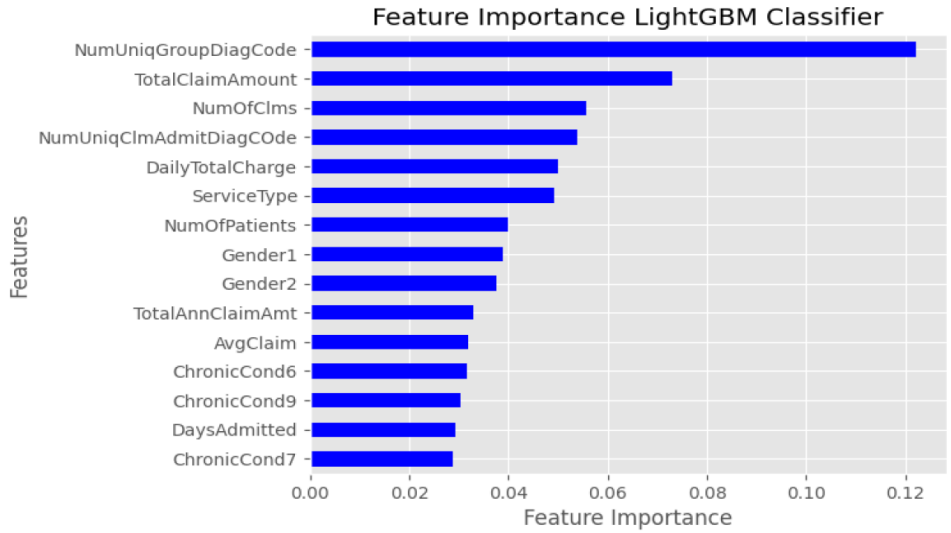


Figure 18: LightGBM Feature Importance

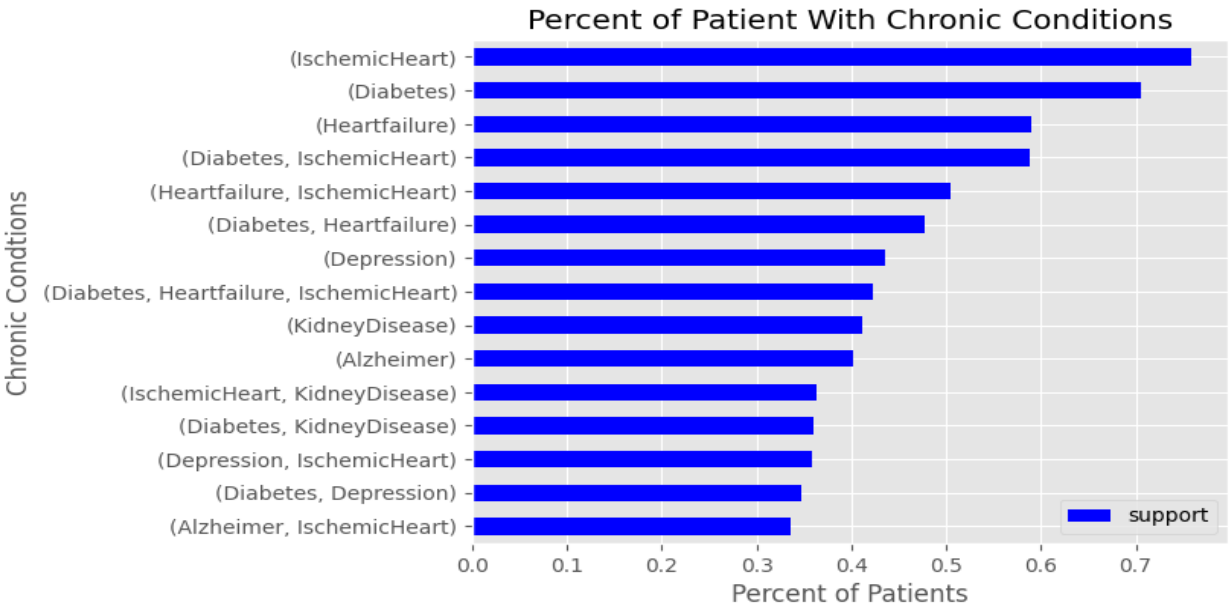
Association Rule Mining on Common Chronic Patterns

- Why `NumUniqGroupDiagCode` is selected as the most important feature ?
- DRG is a way of classifying patients under a particular group by the level of medical resource they use. Each DRG has a payment weight assigned to it. A difference between the assigned DRG categories could double or triple the cost. This is how `numDiffGroupDiagCode` affects the strength of `DailyTotalCharge` and other financial features.
- There are two provider frauds due to this : upcoding and unbundling. Hence, we need to determine what frequent patterns of chronic conditions might be associated with fraudulent providers using the whole data.
- One approach is to apply association rule mining using the Apriori algorithm.
- Items set $I = \{i_1, \ldots, i_m\}$, transactions set $T = \{t_1, \ldots, t_n\}$, $t_i \subseteq I$ and is unique.
- The association rule $X \to Y$, $(X \cap Y = \emptyset)$, is a directed rule between two itemsets. In our context, a patient will likely to have chronic diseases in set Y if she has diseases in set X.
- **Support** measures the occurrence of X in all transactions, $\text{Supp}(X) = \frac{\#\{X\}}{|T|}$. **Confidence** measures the strength of the association rules : the occurrence of Y in all transactions given X, $\text{Conf}(X \to Y) = P(Y \mid X)$. In Apriori algorithm, we need to set minimum support and confidence for mining frequent associations.
- Apriori algorithm in one word : if an itemset is frequent, then all of its subsets are frequent ; if an itemset is not frequent, then all of its supsets are not frequent. **Join :** generates all possible k-itemsets (itemsets with cardinality k). **Pruning :** removes the infrequent candidates according to the minimum support

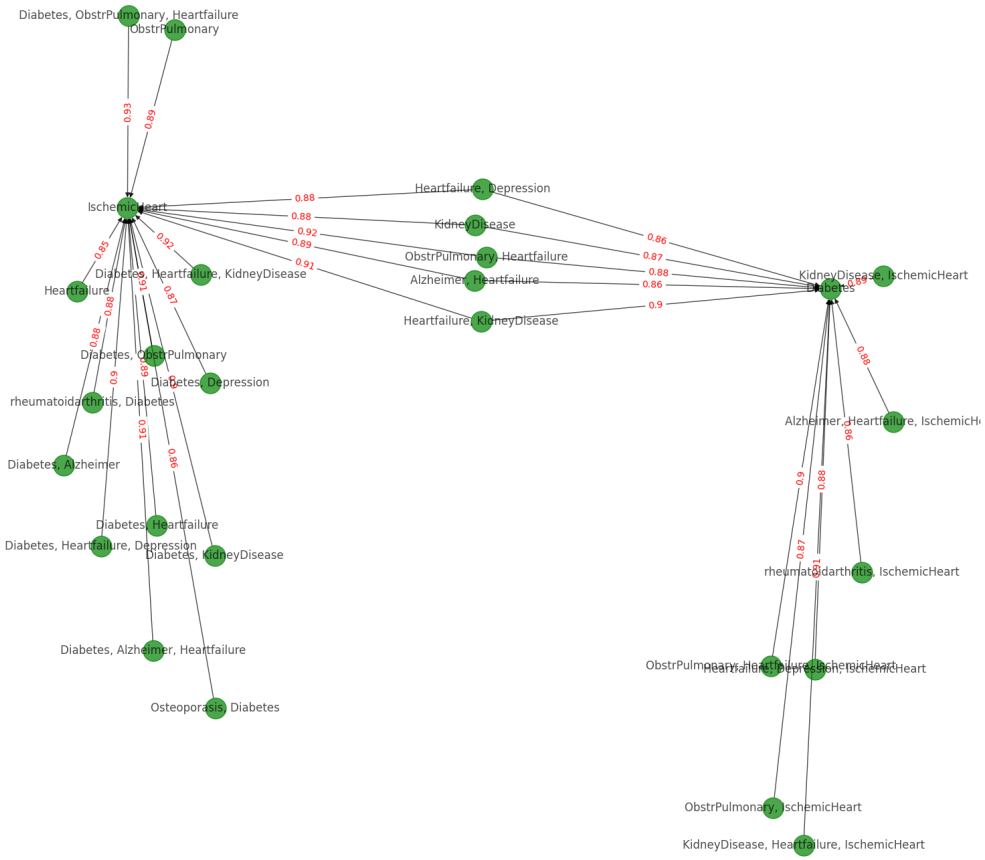
Association Rule Mining Cont'd

Two steps :

- Using the Apriori algorithm, determine all the frequent k-itemsets satisfying the minimum support, which is called frequent itemset X. In our case, we take the threshold to be 0.2.
- Extract association rules with high confidence level, which are called strong rules. For all $A \subsetneq \emptyset$, let $B = X \setminus A$. If Conf$(X \to Y) \geq$ min-Conf, output $A \to B$. In our case, we take the top 30 rules.



Associate Rule Mining Cont'd



Interpretations ?

Among fraudulent providers, 76% of the patients have diabetes, 71% have ischemic heart, and only 59% have heart failure. Patients who have the most common chronic diseases in the healthcare system are susceptible to billing fraud !

Network Analysis

1. Network analysis on provider-patient and provider-doctor at county level :
Plotting **undirected** graphs :

- The weight of the edges are determined by the total reimbursement from all claims filed between the two actors.
- The size of the node is proportional to the strength, the sum of weights of the edges connecting to that particular node.

Below are (bipartite-projected) provider-patient and provider-doctor networks in NY, County # 200 :

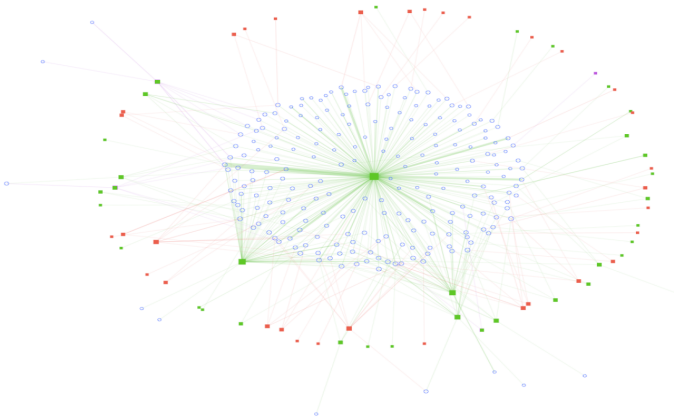


Figure 1: Provider-Patient Network for NY, County # 200

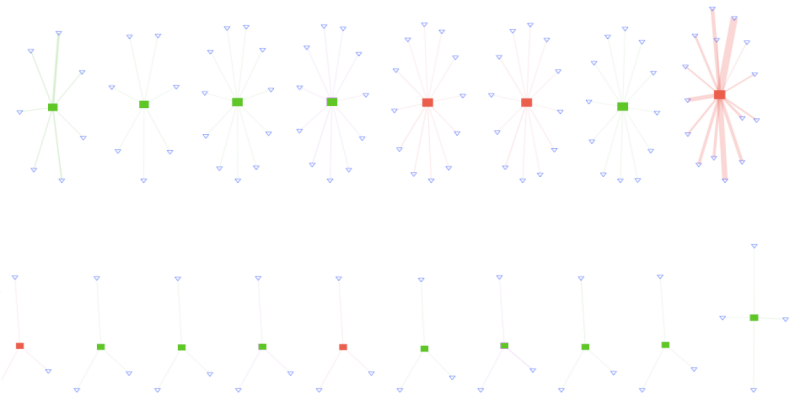


Figure 2: Provider-Doctor Network for NY, County # 200

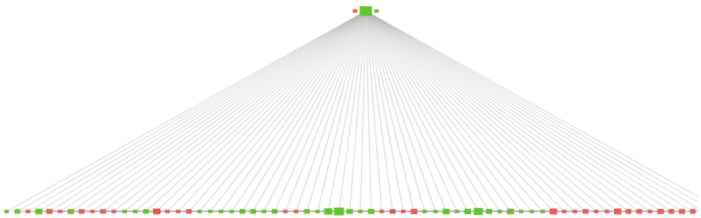


Figure 3: Projected Provider-Patient Network for NY, County # 200

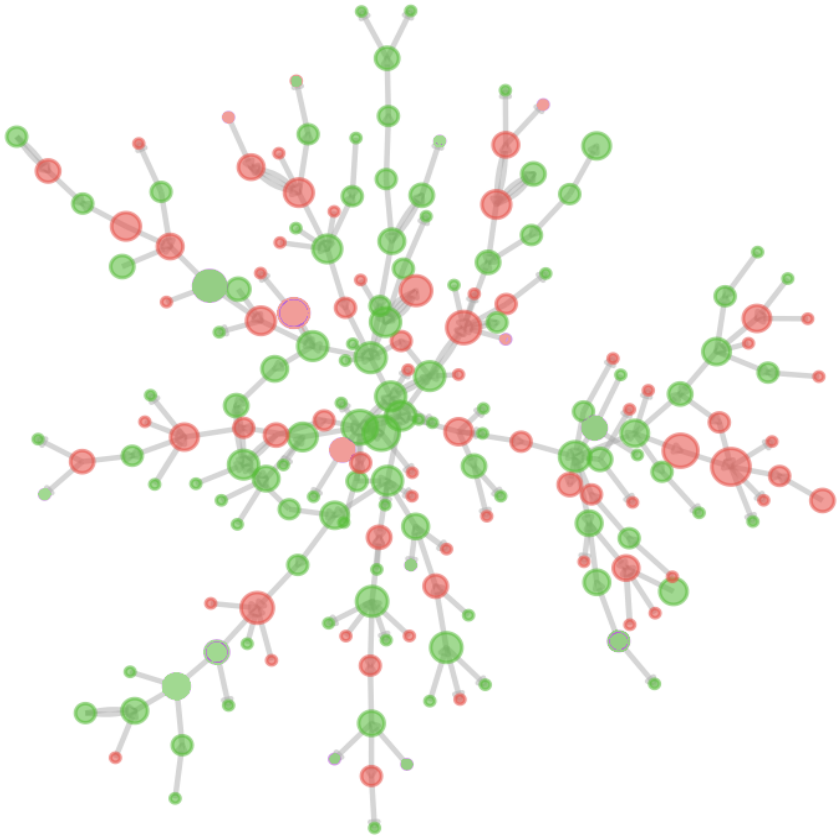
Figure 4: Projected Provider-Doctor Network for NY, County # 200

Network Analysis Cont'd : Detect Duplicated Billings between Providers

2. Network analysis on provider-provider at state level :
Plotting **directed** graphs :

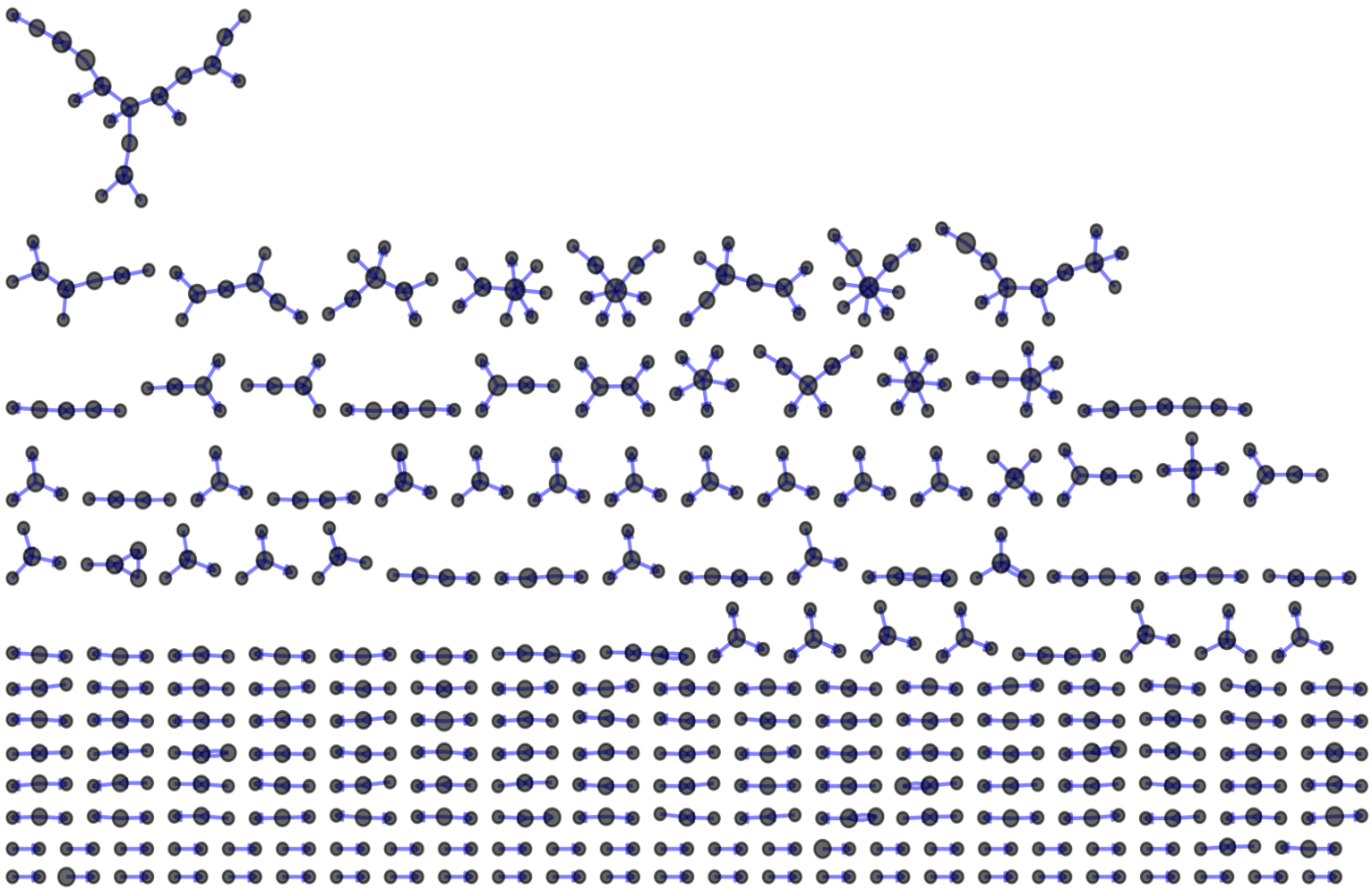
- Identify duplicated claims by finding claims with the same patient having the same three diagnosis codes (`ClmDiagnosisCode`)
- Filter out null entries which likely represented claims associated with patient visits with no diagnosis or procedure performed.

Below is the largest duplicated network of providers found in the data, corresponds to NY-NJ provider network :



Network Analysis Cont'd : Detect Duplicated Billings between Doctors

3. Network analysis on doctor-doctor at state level :
We can use a similar approach for plotting provider-provider networks.
Below is a showcase of portion of duplicated network of doctors found in the data, corresponds to NY doctor network :



Conclusion & Future Work

Based on the three methods we used, we can have a comprehensive framework on characteristics of fraud providers and patients susceptible to fraud :

Patients :

- Aged between 65 and 82, with high number of claims filed.
- With ischemic heart failure and diabetes in their chronic profile, with high number of chronic conditions.
- Received high reimbursement amounts and paid high deductibles, thus with high claim amount.

Providers :

- With high total reimbursements filed in provider-patient network.
- With large networks of duplicated claims between providers, and between doctors.
- With both inpatient and outpatient services.
- With a higher number of DRGs and Claim Admit Diagnosis Code listed on claims.

Recommendations :

- Stricter regulations of assigning diagnosis codes to mitigate risks of upbundling and upcoding.
- Investigation into fraudulent networks, especially on big provider networks and shared patients to mitigate risks of duplicated billing.
- Pay attention to old patients with ischemic heart failure and diabetes

Future Directions :

- Need domain knowledge on DRG assignments to better study upcoding and unbundling.
- Based on network analysis, we can construct graph features and combine them with traditional ML algorithms to boost in prediction accuracy, or use GNNs to learn the graph structure to highlight fraud providers in the network.