The Effect of Enhancing Medicare Claims with OSINT on the Community Membership of Fraud Rings

by

Michelle Espinoza

Kateryna Nesvit, Ph.D., Dissertation Chair

Soumya Sivakumar, Ph.D., Dissertation Internal Committee Member

Daniel Himmelstein, Ph.D., Dissertation External Committee Member

A Dissertation Presented in Partial Fulfillment

of the Requirements for the Degree of

Doctorate of Business Administration

Marymount University

December 2025

MARYMOUNT UNIVERSITY

Dissertation Approval

DBA in Business Administration
School of Business
Marymount University

We hereby approve the dissertation of

Michelle Espinoza

Candidate for the Doctor of Business Administration in Business Intelligence

Date:

Committee Members:

12/10/2025

DocuSigned by:
F8F4790496774F8...
Chair Signature
Name: Kateryna Nesvit, Ph.D.

12/10/2025

DocuSigned by:
AB31075207D74B9...
Internal Member Signature
Name: Soumya Sivakumar, Ph.D.

12/10/2025

DocuSigned by:
EC08EA0A47E8413...
External Member Signature
Name: Daniel Himmelstein, Ph.D.

Signed by:
Accepted by
428E463BBF98456...
Anne Magro, Ph.D.
Dean of Marymount University College of
Business, Innovation, Leadership and Technology

TABLE OF CONTENTS

ABBREVIATIONS

Abbreviation	**Definition**
CMS	Center for Medicare and Medicaid Services
DME	Durable Medical Equipment
GAO	Government Accountability Office
KNN	K-nearest neighbors
MAC	Medicare Administrative Contractor
OIG	Office of Inspector General
OSINT	Open-source intelligence

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

Introduction

Traditional investigative methods are often constrained by siloed data, fragmented information, and resource scarcity, making them ill-equipped to detect sophisticated fraud networks. In this context, network analysis augmented with Open-Source Intelligence (OSINT) offers a promising path forward, enabling the detection of latent connections that would otherwise remain obscured. This dissertation tests whether augmenting Medicare provider claims graphs with Open-Source Intelligence (OSINT) can improve the clustering of confirmed fraud rings using graph-based community detection. Specifically, the study evaluates whether integrating external attributes such as shared addresses, business registration ties, or corporate ownership metadata will increase the likelihood that fraud ring members are clustered together.

In a sea of Medicare claims, where a single provider node may have thousands of transactional or geographic connections, it remains unclear whether adding even a known and confirmed tie, such as a shared corporate officer or physical address, will meaningfully influence cluster assignment. The structural density of these networks raises a key empirical question: Can the deliberate insertion of verified relationships nudge nodes into the same detected communities, or will these links be overwhelmed by the background noise of legitimate but voluminous interactions?

Rather than surveying all possible fraud indicators, this study narrows its scope to three key features commonly associated with Medicare fraud rings:

(1) shared addresses,

(2) overlapping corporate entities or ownership, and

(3) digital traces that reveal operational links (e.g., reused business names, phone numbers, or registrant data).

These were selected based on both prior research and their operational detectability through OSINT.

By establishing these boundaries, the study avoids scope creep and focuses on the evaluative power of targeted graph enrichment in clustering accuracy using the Leiden community detection algorithm. This work hypothesizes that even modest OSINT integration, if strategically aligned, can uncover otherwise obscured community structures.

Medicare fraud rings represent a sophisticated and evolving threat, distinguished from lone actors by their coordination, adaptability, and use of organizational structures to mask illicit activity. Despite their impact, the detection and disruption of such networks remain heavily constrained by investigative bandwidth and the overwhelming scale of available data. The Office of the Inspector General (OIG) continues to decline promising investigations due to resource limitations, with a staggering 15:1 ratio of open Medicare fraud cases to prosecutions (GAO, 2024).

This study extends and applies two key theoretical frameworks: information asymmetry and network homophily. Information asymmetry arises when investigators either lack sufficient data or are inundated by fragmented, unstructured sources, impairing their ability to recognize organized patterns (Akerlof, 1970). Meanwhile, network homophily theory posits that entities with shared attributes such as ownership structures, service types, or billing behaviors, naturally cluster together. However, in the case of fraud rings, this natural clustering is often intentionally obfuscated, creating a paradox: bad actors seek both to collaborate and to mask their proximity.

As a result, clusters that should theoretically exist are artificially fragmented or disguised (Barone & Coscia, 2018; Kurshan & Shen, 2020; Wang et al., 2024).

This obfuscation intensifies the problem of information asymmetry. The central proposition of this study is that strategically integrating OSINT can reduce this asymmetry, by enriching the graph with external attributes that reveal hidden or masked relationships, thereby making the underlying homophily detectable once again.

Although Medicare fraud is rarely classified as cyber-enabled crime, it increasingly relies on the same digital infrastructure such as falsified identities, remote operations, and cybercrime-as-a-service ecosystems that powers transnational fraud schemes. The ecosystem of cybercrime-as-a-service and platform enterprises has significantly enhanced the capability of fraudsters to collaborate internationally to execute illicit activities in jurisdictions where, at times, none of the perpetrators are physically present. Case examples involve a durable medical equipment (DME) company owner who sold his business online to an individual with a counterfeit driver's license. The fictitious owner is now using the company as part of a suspected multimillion dollar Medicare fraud ring (McNicholas & Randolph, 2024). In another major case, officials disrupted a $1 billion fraud ring that involved US DME companies contracting with offshore centers, who in turn used online advertising to recruit US physicians to participate in an elaborate kickback scheme (Federal Bureau of Investigation, 2019). Both cases demonstrate how fraudsters exploit online platforms to scale their schemes with near impunity.

Prior work has shown that bipartite and attribute-enriched networks can expose previously hidden fraud patterns and associations (O'Malley et al., 2023; Wang et al., 2024). Extending and applying this foundation, the present study tests whether OSINT integration into

graph structures can further enhance community detection of fraud rings. While it may seem counterintuitive to address information overload by adding more data, the hypothesis is that targeted OSINT enrichment can improve the signal-to-noise ratio, surfacing meaningful associations that are not evident in claims data alone.

Through a quasi-experimental design, this study compares community detection results between baseline graphs and OSINT-augmented graphs using the F1-score to measure alignment with known fraud labels. In doing so, it aims to assess whether OSINT can function not only as an investigative tool but as a graph augmentation strategy that meaningfully improves fraud ring detection.

Background of the Study

The Office of the Inspector General (OIG) is responsible for Medicare oversight. Despite being responsible for more than 100 programs through Health and Human Services (HHS) 80% of the OIG's budget is spent on Medicare fraud investigation and enforcement. Despite this focus, the OIG acknowledges an ongoing inability to keep pace with the volume of fraud. Each year, the agency declines an estimated 300–400 viable criminal cases due to resource constraints—excluding the many more reports that go uninvestigated.

By one OIG inspector's own estimate, 90% of Florida's durable medical supply companies that are billing Medicare, are fraudulent or non-existent (Wilson, 2023). In 2022-2023, 18,129 durable medical supply companies in Florida billed Medicare. If the inspector's estimate is accurate, this estimate suggests that over 16,000 providers in a single state may be engaging in fraudulent activity unchecked.

According to the National Healthcare Antifraud Association, taxpayers lose more than $100 Billion per year to Medicare Fraud and say that is a conservative estimate. The government spent $900 Billion on Medicare in 2022. During the OIG's routine auditing of claim reimbursements, they found overbilling or unjustifiable charges in 65-70% of their audit samples, further underscoring the scope and pervasiveness of fraudulent activity.

Against this backdrop, Open-Source Intelligence (OSINT) has gained traction as a powerful tool for augmenting investigative processes. Defined as the collection and analysis of publicly available information—including data from websites, social media, and public records—OSINT has been supercharged by advances in big data and AI. These technologies now enable real-time processing, contextual linkage, and predictive insights at a scale previously unattainable (Kumar et al., 2024).

In the context of Medicare fraud, graphs—modeled as nodes (entities) and edges (relationships)—offer a methodologically rich structure to represent complex networks of interaction. These graphs can incorporate multiple edge types, weighted relationships, and layered attributes. OSINT data can enhance these structures by exposing hidden relationships, enriching node properties, and revealing cross-entity affiliations that are not apparent from claims data alone.

This study examines a critical and unresolved theoretical tension: Do Medicare fraud actors naturally cluster in networks, or do they remain distant by design? According to network homophily theory, entities with similar characteristics—such as billing behaviors, services rendered, or corporate affiliations—should gravitate toward one another and form detectable communities. Yet in practice, many fraud rings rely on intentional obfuscation tactics, such as

straw ownership, third-party recruiters, or cross-jurisdictional entities, to mask those very relationships.

From this tension arises a third explanatory lens: information asymmetry theory. It proposes that both patterns may be true—fraudsters may collaborate and share resources, but unless the investigative lens is sufficiently informed, those ties remain hidden. In this context, whether fraud actors are detected within the same cluster depends not just on their behavior, but on the degree of information asymmetry between them and the entities tasked with oversight. This study uses OSINT-enriched graphs as an intervention aimed at reducing that asymmetry and testing whether more complete information results in better-aligned community detection outcomes.

Problem Statement

Fraud cost companies an average of 5% in annual revenue in 2022, with total global losses estimated in the trillions (Association of Certified Fraud Examiners, 2022). Among the most damaging and elusive forms of fraud are those committed by organized rings. These groups act in concert to exploit systemic gaps, often adapting quickly to evade traditional detection mechanisms. Their coordinated behavior makes them especially challenging to detect using siloed or transactional-level data alone.

The general business problem is that detecting fraud rings, particularly in complex, high-volume domains like healthcare—is notoriously difficult. Current detection methods often prove ineffective against sophisticated schemes, and fraudsters continuously evolve their tactics to stay ahead of traditional rule-based or machine learning systems. The specific business problem is

that fraud detection systems frequently rely on internal domain knowledge and proprietary data, limiting the development and testing of more generalizable or graph-based approaches (Grover et al., 2022). Publicly available datasets are scarce, and even when graphs are used, they are often constructed from a single data source, limiting visibility into cross-entity relationships. Healthcare provider networks contain intricate community structures that are often incompletely captured when analyzed through single-source data (Ran et al., 2024). Understanding how these communities form and interconnect requires addressing information asymmetries inherent in limited datasets. This study examines how augmenting network graphs with Open Source Intelligence (OSINT) affects community detection and entity clustering within Medicare provider claims data.

Purpose of the Study

The purpose of this quasi-experimental quantitative study is to assess whether strategically integrating Open-Source Intelligence (OSINT) into Medicare provider graphs improves the detection of known fraud rings. The study focuses specifically on three high-signal indicators of fraudulent community membership:

(1) shared business or mailing addresses,

(2) corporate ownership or registration linkages, and

(3) digital identifiers such as phone number or business name reuse.

These OSINT-derived attributes are used to augment provider graphs that were originally constructed from Medicare Part B and Part D claims data.

Through a comparative analysis using the Leiden community detection algorithm, this research evaluates whether OSINT augmentation improves clustering performance as measured by the F1-score—a standard metric used for binary classification models. By operationalizing OSINT in a bounded, repeatable way, the study seeks to contribute to both investigative practice and the broader understanding of graph enrichment strategies in fraud analytics.

Definition of Terms

Medicare Part B Claims- claims for doctors' services, outpatient hospital care, ambulance services, medical supplies and equipment

Medicare Part D Claims - claims for durable medical devices such as prosthetic limbs, wheelchairs and medication prescriptions

OSINT - Open-source intelligence produced by collecting, evaluating and analyzing publicly available information with the purpose of answering a specific intelligence question.

Fraud Ring – A group of two or more actors colluding in fraudulent activity for monetary gain.

Significance of the Study

This study contributes to the advancement of fraud analytics by exploring how fraud rings appear in the Medicare claims data and how augmenting Medicare provider graphs with Open-Source Intelligence (OSINT) can improve the identification of hidden or loosely connected fraud rings. Specifically, it addresses challenges in modeling real-world, multi-entity networks where relationships are often obscured or misrepresented due to incomplete data.

The findings offer practical significance for investigators and analysts working in healthcare fraud, enabling more scalable and effective graph enrichment techniques. From a theoretical standpoint, the study advances understanding of how data fusion impacts community detection and the interpretability of network-based models in adversarial settings. The broader contribution is a step toward more systemic, resilient approaches to fraud detection in critical public programs.

Research Questions

The proposed research questions for this dissertation are:

1. Does augmenting Medicare provider graphs with three specific OSINT-derived indicators—shared address, corporate registration links, and digital trace attributes—improve the detection of confirmed fraud rings, as measured by the F1-score, when compared to baseline graphs without OSINT?
2. Can FastRP embeddings applied to OSINT-augmented graphs preserve comparable F1-score performance while reducing graph complexity?

Theoretical Perspective

A theory in quantitative research is a set of interrelated constructs, definitions, and propositions that present a systematic view of phenomena by specifying relationships between variables with the purpose of explaining phenomena (Schneider & Kerlinger, 1979). The relationships exist in the form of propositions or hypotheses (Creswell & Creswell, 2022).

The theoretical framework for this study draws from three interrelated traditions: Asymmetric Information Theory and its consequence, Moral Hazard, from Economics, and

Graph Theory, from Mathematics. Together, these theories provide a systematic lens for understanding the challenges of detecting fraud rings within the Medicare system.

Asymmetric Information Theory, borrowed from Economics, posits that in transactions, one party possesses more or better information than the other, leading to an imbalance that can be exploited (Akerlof, 1970). In the context of Medicare fraud, this theory is particularly relevant as healthcare providers often have superior knowledge about medicine over those who process the Medicare billing and payments. Additionally, providers can bill multiple insurance companies, and each company only sees a portion of a provider's full billing activity. This information asymmetry creates opportunities for fraudulent behavior, as providers can manipulate billing processes and exploit loopholes without immediate detection. The theory suggests that the lack of transparency in healthcare transactions can lead to increased fraud, as patients and even regulatory bodies may not fully understand the services rendered or the legitimacy of the claims submitted.

Moral hazard further intensifies this vulnerability. When actors believe they are unlikely to bear the consequences of risky behavior, they may engage in greater levels of misconduct. In the Medicare context, resource limitations at the Office of the Inspector General (OIG) constrain the investigation and prosecution of fraud, often delaying enforcement by years. This low-certainty, low-celerity environment reduces deterrence, creating a high-reward, low-risk ecosystem attractive to bad actors.

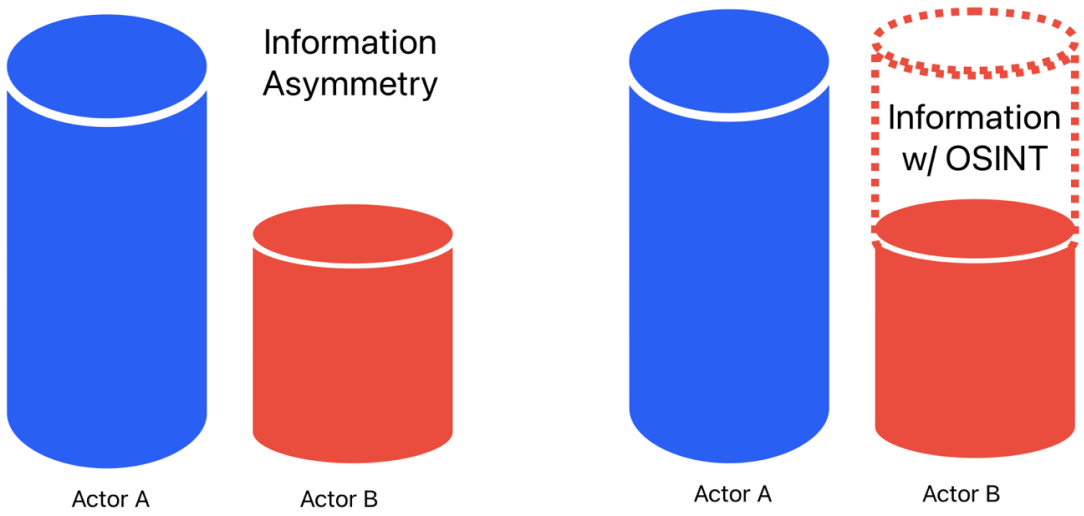
Graph Theory offers a mathematical framework for modeling and analyzing relationships between entities—in this case, providers, addresses, medical procedures, medications, and equipment. Graph-based approaches allow for the visualization and detection of complex

network structures, including the identification of communities that may represent collusive fraud rings.

However, organized fraud networks present an additional challenge: the paradox of corrupt networks. According to network homophily theory, entities with shared attributes naturally tend to cluster together. Yet, fraudulent actors simultaneously collaborate and seek to obscure their ties to evade detection. This creates intentional fragmentation within the network structure, masking what would otherwise be visible communities (Barone & Coscia, 2018; Wang et al., 2024). Thus, the natural homophily expected among collaborating providers becomes deliberately obfuscated.

This study proposes that augmenting Medicare claims data with strategically targeted Open-Source Intelligence (OSINT) can reduce information asymmetry, reveal hidden homophily, and thus counteract the obfuscation tactics employed by fraud rings. OSINT augmentation aims to increase the informational content of the network, making latent relationships more detectable without overwhelming the system with noise. Through this framework, the study tests whether reducing information asymmetry via OSINT integration can improve the accuracy of fraud ring detection as measured by community detection performance.

Figure 1-1
Information Asymmetry Before and After OSINT



Extending and applying these theoretical constructs, this study applies graph-based community detection methods to uncover fraud networks within Medicare claims data. Recognizing the limitations of claims data alone, open-source intelligence (OSINT) is incorporated to augment the informational content of provider networks. This augmentation is theorized to reduce information asymmetry, mitigate the effects of moral hazard, and improve the accuracy of fraud ring detection. By enriching graph attributes with external data, the analysis aims to reveal hidden structures and patterns of collusion that would otherwise remain undetected. Together, these strategies form the conceptual basis for the study's methodology and provide a systematic approach to investigating the dynamics of fraud networks under conditions of incomplete information.

Rather than uncovering ***unknown*** fraud, this study focuses on evaluating how augmenting Medicare claims data with open-source intelligence (OSINT) affects the detection of

confirmed fraud networks. Specifically, it tests whether adding external, publicly available information reduces information asymmetry and improves the accuracy of community detection with the Leiden algorithm. By comparing community membership alignment before and after OSINT augmentation—measured against known fraud ring labels—this research assesses whether supplementary information enhances the graph's ability to accurately cluster related fraudulent actors. In doing so, it shifts the challenge from the discovery of unknown entities to the refinement of analytical methods for better identification and characterization of known fraud schemes. This approach lays the foundation for the literature review, which synthesizes prior research on OSINT, graph-based methods, network homophily theory, and healthcare fraud.

Limitations of the study

A major limitation of the study is that confirmation of fraud ring membership relies on conviction and/or official exclusion from the Medicare program. With thousands of unprosecuted and undetected cases amassing each year, the Medicare exclusionary database is not a complete representation of all fraudulent actors in the Medicare claims data set.

Another limitation of the study is that it relies on publicly available data and does not account for the possibility of falsified medical credentials, false identities, straw ownership of shell corporations, or unnamed defendants some of whom may still be working with law enforcement as part of their plea agreement. These factors create significant gaps in the understanding of the full scope of fraudulent activities, making it difficult to draw comprehensive conclusions about the effectiveness of augmenting Medicare claims with OSINT. A final limitation of this study is the reliance on historical data, which may not accurately reflect current trends in fraud schemes or the evolving tactics employed by fraudulent actors.

CHAPTER 2. LITERATURE REVIEW

This literature review synthesizes research on Medicare fraud, graph-based analytics, and the use of Open-Source Intelligence (OSINT) in network analysis. It focuses on three high-signal OSINT indicators that guide this study:

(1) shared addresses,

(2) corporate registration ties, and

(3) digital identifiers such as reused phone numbers or business names.

These indicators anchor the review, providing boundaries that prevent sprawl while connecting each strand of literature directly to the research questions.

The chapter proceeds in four stages. First, it reviews the persistence and scale of Medicare fraud, particularly the systemic weaknesses that are exploited. Second, it examines how claims processing structures constrain detection and contribute to information asymmetry. Third, it evaluates graph-based approaches, particularly community detection, FastRP embeddings, and the Leiden algorithm—as methodological responses to these challenges. Finally, it grounds the study in network homophily and information-asymmetry theory and reviews prior OSINT applications.

Combined, these topics establish the empirical and theoretical gaps that this study addresses in testing whether targeted OSINT enrichment can improve the clustering of known fraud rings in Medicare provider graphs.

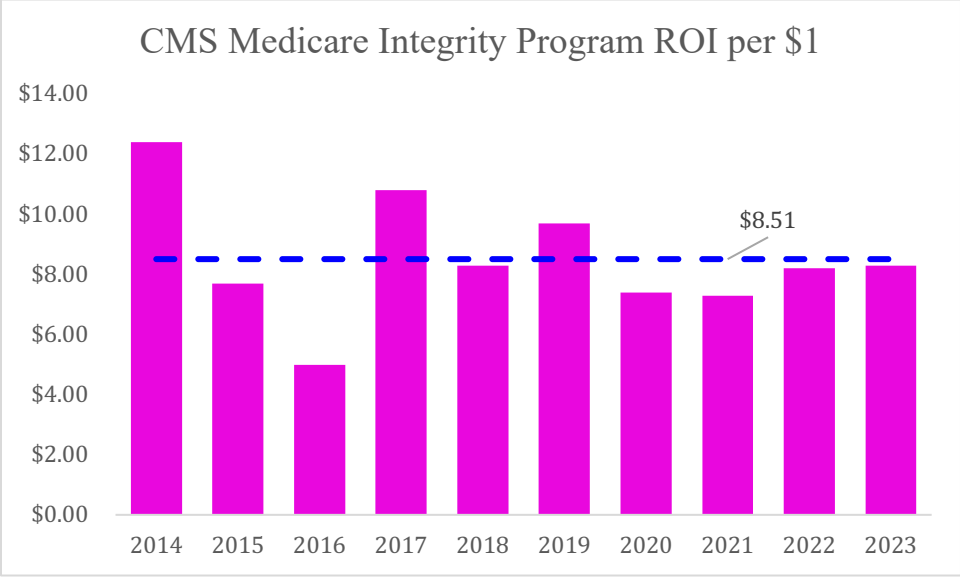
Medicare Fraud

Medicare fraud represents a persistent and costly challenge for federal healthcare programs, consuming billions in taxpayer funds each year and stretching investigative capacity

to its limits, making traditional methods insufficient and motivating the need for OSINT-based graph enrichment.

Fraud rings exploit systemic weaknesses in provider enrollment, claims processing, and oversight coordination, making them difficult to detect using standard transactional reviews. During the 2013–2014 fiscal years, the Department of Health and Human Services (DHHS) and Centers for Medicare and Medicaid Services (CMS) reported a return on investment (ROI) of $12.40 for every dollar spent combating fraud, which dropped to $7.70 by 2015. Over the lifespan of the Health Care Fraud and Abuse Control (HCFAC) Program, joint efforts by the DHHS and Department of Justice (DOJ) returned $27.8 billion to the Medicare Trust Fund. Prevention and enforcement are split across multiple agencies. CMS' Medicare & Medicaid Integrity Program annual report to congress reported an average of $8.51 per $1 spent between 2014 and 2023 (Fig. 2.1).

Figure 2-1
CMS Medicare and Medicaid ROI by Year



During this period, the government's strategy shifted from a reactive "pay-and-chase" model to a proactive fraud prevention approach. CMS emphasized stricter provider enrollment safeguards to ensure only legitimate providers could bill Medicare. Real-time data analysis was introduced to expedite the identification, arrest, and prosecution of fraudsters, bypassing slower traditional methods such as subpoenas and manual account analysis.

Between 2009 and 2014, the Health Care Fraud Prevention and Enforcement Action Team (HEAT) charged 2,097 defendants across 963 cases, resulting in 1,197 convictions with an average prison sentence of 47 months. In 2014 alone, 924 new criminal investigations were initiated, 496 cases were filed, and 734 convictions were secured. Civil enforcement efforts also yielded significant results: the False Claims Act obtained $2.3 billion in settlements and judgments from healthcare fraud cases in 2014, contributing to a total civil recovery of $15.2 billion between 2009 and 2014. Fig. 2 shows the total Medicaid investigations open at the end of each year. Similarly comprehensive annual statistics are not available for Medicare, but Medicaid spending is approximately 87% of Medicare spending. According to one convicted fraudster, "It's just so easy, it's unbelievable…They get caught, they get out, and they'll do it all over again. I don't think the government can keep up. People keep on. They're not gonna stop" (Zamost & Brewer, 2023). His statements about repeat offenders echo my observations from reviewing individual cases.

Figure 2-2
Open Investigations by Year and Type

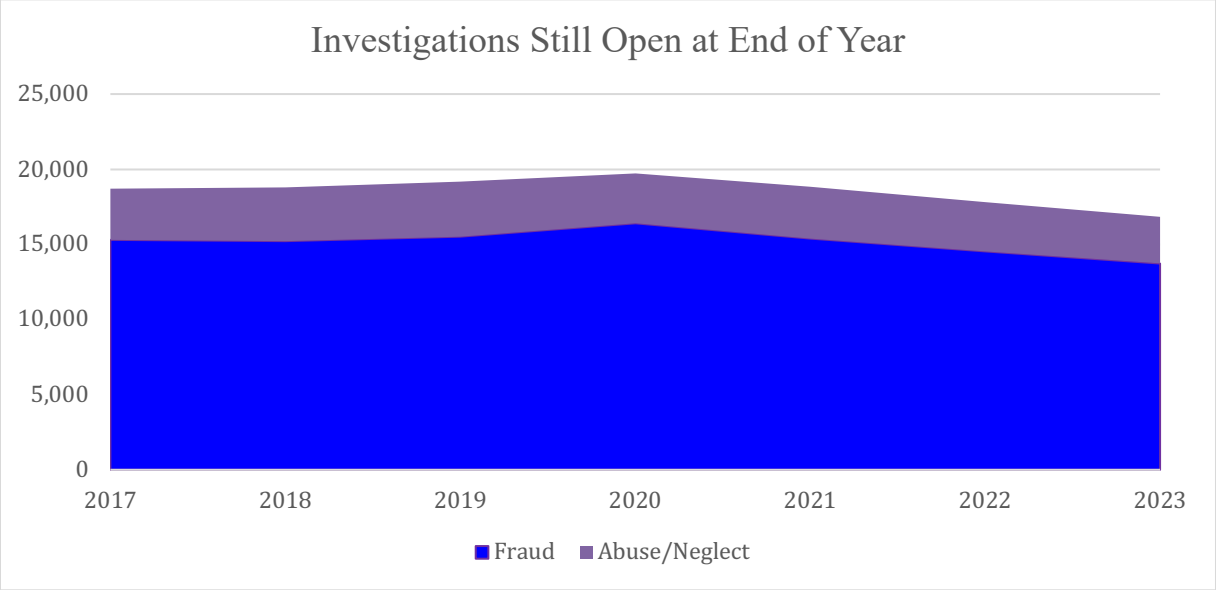
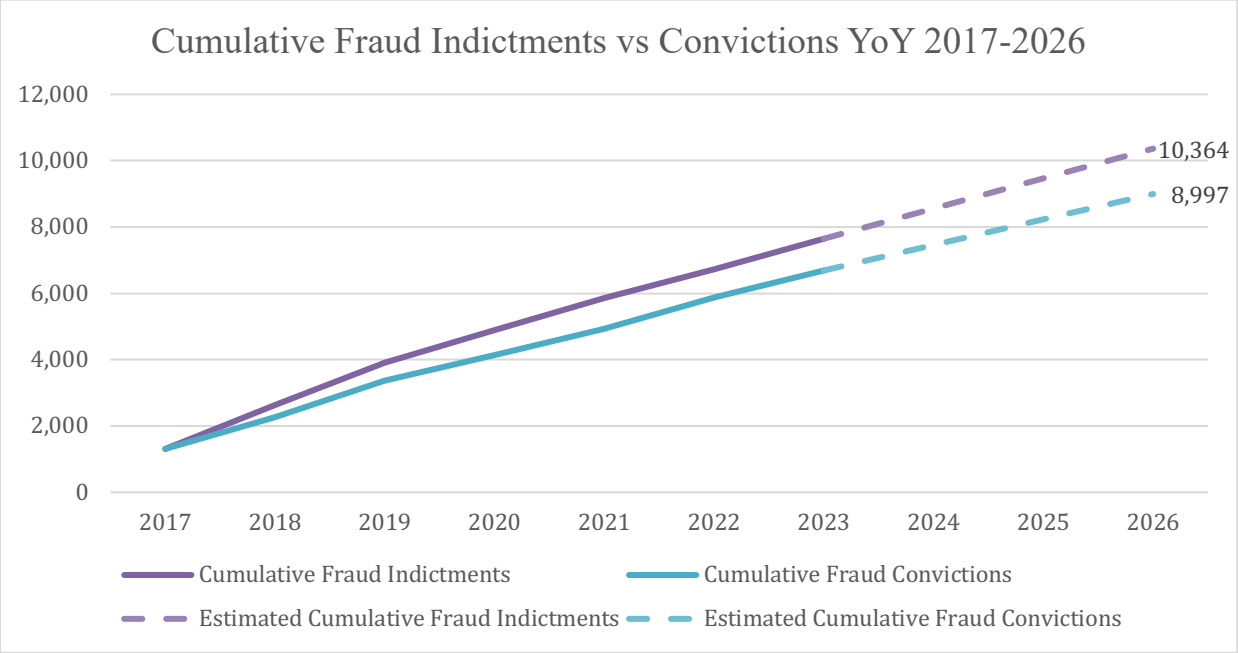


Figure 2-3

Cumulative Medicare fraud indictments vs convictions year over year



CMS also undertook major provider revalidation efforts. In 2015, all 1.5 million Medicare suppliers were required to revalidate their eligibility under enhanced screening

standards, resulting in 450,000 deactivations and 27,000 revocations. Deactivated providers could reapply, while revoked providers faced a one- to three-year bar from the program. Key fraud "hotspots" identified during this period included Miami, Chicago, Dallas, Houston, Detroit, and Philadelphia, as well as parts of New Jersey.

Despite aggressive enforcement, the 2022 HCFAC Annual Report noted that ROI had declined to $2.90 per dollar spent, attributing the decrease to multiple factors, including changes in case complexity and enforcement strategies.

Medicare Claims Processing

The structure and pace of Medicare claims processing create conditions that favor fraud by prioritizing payment speed over comprehensive review. This creates systemic blind spots that claims-only methods cannot overcome; OSINT attributes are necessary to close those gaps. Medicare claims must be processed according to guidelines outlined in Chapter 26 of the Medicare Claims Processing Manual. Under the Administrative Simplification Compliance Act (ASCA), most claims must be submitted electronically, with only limited exceptions.

Medicare claims must be adjudicated within 30 days, and after 90 days, interest must be paid on any outstanding amounts. Graph databases facilitate faster lookup of related transactions associated with claims. The requirement for rapid claims processing, combined with the high volume of claims, creates significant challenges for fraud detection. Certain claim types, such as Hospice and Home Health services, are reimbursed almost immediately after submission, leaving little time for thorough review.

Although Medicare is a federally funded program, claims processing is decentralized through private contractors. Originally managed by Part A fiscal intermediaries and Part B carriers, the 2003 Medicare Prescription Drug, Improvement, and Modernization Act (MMA) consolidated these roles under Medicare Administrative Contractors (MACs). Today, 12 A/B MACs and 4 Durable Medical Equipment (DME) MACs administer claims for nearly 54% of 35 million Medicare beneficiaries.

MACs are responsible for enrolling providers, processing claims, issuing payments, and ensuring compliance with Medicare rules within their designated jurisdictions. This fragmentation means no single contractor sees the full relational footprint of a provider or ring; actors can distribute activity across jurisdictions to appear unconnected in any one dataset. The next section addresses the measurement problem: why official figures provide only a lower bound and why OSINT is needed to surface hidden relationships.

Figure 2-4
MAC Jurisdictions for DMEs

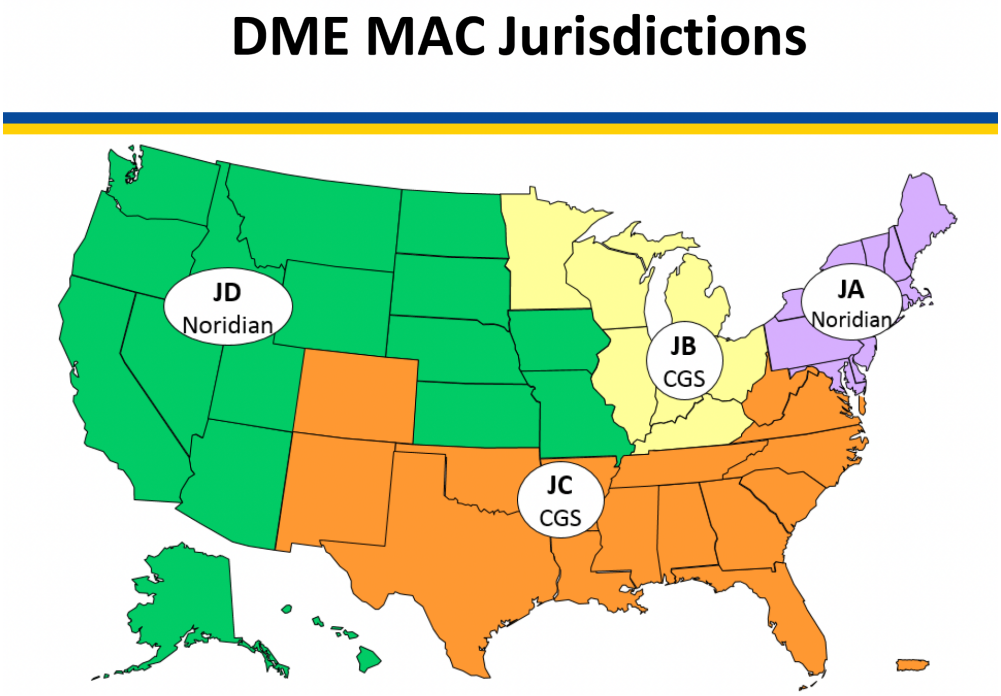
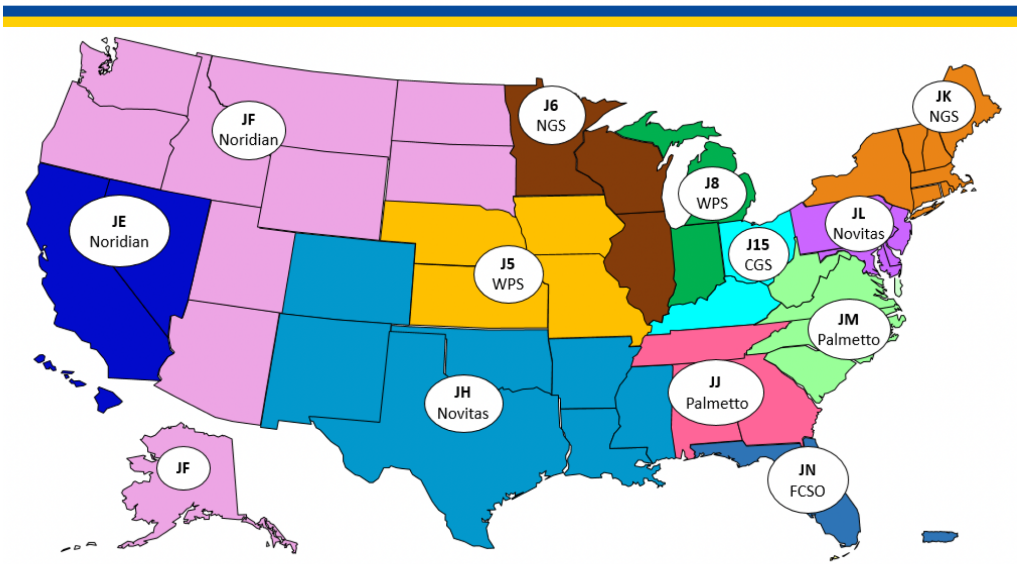


Figure 2-5
MAC Jurisdictions for Medicare Parts A&B Claims



Estimating Fraud in Medicare

Official statistics reflect prosecuted or audited cases and thus undercount the true prevalence of fraud. Proxy methods help, but remain constrained by claims-only data which reinforces the need for OSINT enrichment. This asymmetry means that current estimates often serve as lower bounds rather than accurate measures of fraud prevalence. Researchers have used creative proxies such as regional spending outliers, proximity to Strike Force offices, peer benchmarking, sequence outliers, and validation against DOJ releases to identify suspicious clusters. These advance measurement but still miss high-risk ties absent from claims files (shared addresses, common ownership, repeated digital identifiers). Without those attributes, collusive providers can look unrelated. O'Malley et al. (2023) applied network analysis techniques to examine the diffusion of fraud in Medicare home health claims, utilizing proxy measures such as regional expenditure outliers and proximity to Strike Force office locations (O'Malley et al., 2023). Similarly, Shekhar et al. (2023) combined outlier statistics, peer-based comparisons, and billing code sequencing to develop an explainable unsupervised learning model for detecting Medicare fraud. Their models were validated against Department of Justice press releases and case studies identifying suspicious providers, helping establish ground-truth labels (Shekhar et al., 2023). These approaches illustrate creative use of available data, but they remain constrained by the information contained within claims systems.

Importantly, the literature shows that many high-risk relationships such as shared physical addresses and common ownership structures activity are absent from these claims-based models. Without external enrichment, estimation methods risk missing clusters of collusive providers who appear unrelated in transactional data alone. This limitation points directly to the

value of targeted OSINT integration. By supplementing claims graphs with verified external attributes, it may be possible to reduce information asymmetry and improve both the accuracy and interpretability of fraud detection models. The next section examines community detection as a method for operationalizing this approach within graph-based Medicare fraud analytics.

Community Detection in Fraud Networks

Community detection is essential for uncovering coordinated Medicare fraud because fraud schemes are more often executed by groups, not isolated actors. Community detection works best when graphs include OSINT-derived edges that increase intra-ring density and separation from benign neighborhoods. Community detection is a critical component of network analysis, particularly for identifying coordinated fraudulent activities within healthcare systems. Traditional methods such as modularity optimization and spectral clustering have been widely used to uncover hidden structures in graphs. However, fraud networks often evolve rapidly and deliberately obscure their connections, limiting the effectiveness of static, structure-only approaches. Claims-only graphs lack the external relationship data (shared addresses, ownership ties, corporate metadata) needed to reveal those groups, limiting detection accuracy and identification.

Studies suggest that augmenting graphs with external data can significantly improve community detection outcomes. For example, Leskovec et al., (2010) demonstrated that integrating additional attribute information into social networks enhanced the identification of meaningful communities (Leskovec et al., 2010). This approach is particularly relevant in fraud detection, where identifying collusion can be modeled as a community detection problem on graphs where nodes are actors and the edges represent common attributes between them

(Masihullah et al., 2022). To address the challenge of fraud detection in large graphs, Massihullah et al. (2022) proposed domain-aware weighted community detection for fraud rings. Inspired by this approach, the present study investigates whether supplementing Medicare claims graphs with open-source intelligence (OSINT) can improve the detection of confirmed fraud rings.

Additionally, this research explores whether node embedding techniques, specifically FastRP embeddings paired with a K-Nearest Neighbor (KNN) edge augmentation strategy based on cosine similarity, can maintain or enhance detection accuracy while reducing graph complexity.

Incorporating targeted OSINT indicators changes graph topology in ways that are meaningful for community detection:

1. Shared addresses add edges that proxy for co-located operations
2. Corporate ownership/registration ties expose control relationships across entities.
3. Digital identifiers surface coordination across otherwise separate providers.

Traditional methods of detecting Medicare fraud often struggle to capture the complexity and interconnectedness of fraudulent schemes. Because these schemes frequently involve multiple coordinated actors, an entity-by-entity analysis risks missing critical patterns of collusion. To address these challenges, this study applies network theory and graph-based methods, which are specifically designed to reveal hidden relationships and community structures within complex systems.

Network Theory

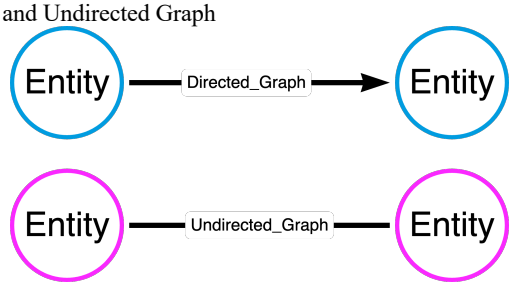
Network theory models Medicare fraud as relationships among entities. Mapping OSINT indicators to edge types can make collusion structurally visible to community detection. Almost all complex systems can be effectively modeled as networks. The study of networks is one of the youngest and most active areas of complexity science, driven less by models and equations and more by the analysis of large, complex datasets. The primary contributions to this field have come from graph theory and computer science. In the context of Medicare fraud detection, this framework enables the representation of providers, facilities, corporate entities, and other actors as nodes, with edges encoding operational, financial, or administrative relationships.

A key advantage of decentralized networks is their robustness to failure; there are no critical or strategic nodes whose removal would cause systemic collapse. This explains why dark web markets often remain resilient even after major busts or seizures (Ouellet et al., 2022; Reis et al., 2023). However, the lack of centrality also results in slow diffusion across networks.

Nodes/Edges in Graphs

Nodes typically represent entities and edges are the relationships between entities. Graphs can be directed or undirected. In a directed graph, edges have a direction (e.g., a one-way street), indicating a flow or transfer from one node to another. In an undirected graph, edges represent bidirectional relationships, like a rope connecting two nodes.

Figure 2-6
Directed and Undirected Graph



Because enriched graphs can grow large, the next section motivates FastRP embeddings as a scalable way to preserve structure while controlling complexity.

FastRP Embeddings

FastRP produces low-dimensional node representations that preserve multi-hop structure efficiently and enable scalable community detection on OSINT-enriched graphs while reducing computational consumption. In graph analysis, the FastRP algorithm provides a scalable and efficient method for generating low-dimensional embeddings while preserving essential structural information. These embeddings maintain high levels of accuracy while significantly reducing computational complexity (Chen et al., 2019). The key question is whether these simplifications preserve detection performance relative to OSINT-augmented, non-embedded graphs.

Leiden Community Detection Methods

Community detection is a central task in network analysis, aimed at identifying clusters or groups of nodes that are more densely connected to each other than to the rest of the network. Among the many algorithms developed for this purpose, the Leiden algorithm stands out for its speed, scalability, and the production of well-connected, high-quality communities (Hairol

Anuar et al., 2024; Traag et al., 2019). The Leiden algorithm is especially suited for large and complex networks, such as those encountered in fraud detection.

In healthcare fraud investigations, community detection plays a pivotal role by uncovering hidden structures within provider networks. Fraud schemes often involve multiple actors—such as physicians, pharmacies, durable medical equipment providers, and shell companies—collaborating across shared addresses, patients, or billing codes. These relationships may be subtle or deliberately obscured, making traditional methods insufficient. Graph-based approaches, particularly those leveraging Leiden, can reveal such structures effectively.

The Leiden algorithm was introduced as an improvement over the widely used Louvain algorithm, which, despite its popularity, suffers from certain limitations—most notably the tendency to produce disconnected or poorly connected communities. The Leiden algorithm addresses this by incorporating an additional refinement phase that ensures each community is internally well-connected. This leads to more robust and interpretable community structures, which are essential when drawing conclusions about the behavior of entities within a network, such as identifying fraud rings or collusive groups.

At its core, the Leiden algorithm works by optimizing a quality function—typically modularity or Constant Potts Model (CPM)—to determine the best partitioning of the network. The algorithm proceeds in iterative phases: it begins by locally moving nodes to improve the quality function, then aggregates the network based on the identified communities, and finally refines the partitioning to ensure connectivity. This iterative approach not only improves performance over Louvain but also reduces the risk of converging on suboptimal or unstable community assignments (Traag et al., 2019).

In fraud detection contexts, the Leiden algorithm is especially useful because fraud networks often display subtle patterns of coordination that may not be apparent through traditional statistical approaches. By leveraging the algorithm's capacity to detect fine-grained structures in large graphs, investigators can uncover clusters of providers or entities that exhibit suspicious co-claiming behaviors, shared addresses, or overlapping organizational affiliations. The high resolution and stability of Leiden's output make it a strong choice for applications that rely on repeatable and interpretable graph-based insights.

The Resolution Limit Problem

A critical consideration in community detection, particularly in fraud detection, is the resolution limit problem (Hairol Anuar et al., 2024; Traag et al., 2019). This phenomenon occurs when an algorithm fails to detect small but significant communities within a large graph, favoring the merger of small clusters into larger communities even when the smaller groups are internally cohesive. This potentially masks collusive rings that are relevant to enforcement. For Medicare fraud, this limitation is particularly acute. Fraud rings are often small in absolute terms, with a handful of providers or companies coordinating billing, easily obscured within tens of thousands of legitimate nodes.

While the Leiden algorithm mitigates this issue better than its predecessors, it does not eliminate it entirely. However, by adjusting the resolution parameter, researchers can fine-tune community sensitivity to capture both macro- and micro-scale patterns. This study explicitly incorporates the tradeoff between fragmentation and detection by adjusting the resolution parameter in each trial and validating partitions against known fraud labels using the F1-score.

By testing OSINT augmentation under varying resolutions, this research evaluates whether additional external edges make small, collusive communities more detectable.

Modularity

Evaluating the quality of community partitions requires metrics that distinguish between clusters formed by chance and those that reflect meaningful structure. Modularity is a common metric to assess the quality of community partitions. Modularity assesses whether the density of edges inside detected communities is greater than what would be expected in a random graph of similar size.

In the context of Medicare fraud detection, modularity is a way to assess whether collusive groups are internally cohesive enough to be distinguished from the broader provider network. A high modularity score indicates strong separation between communities, while a low score suggests that boundaries are less well defined. However, modularity is closely tied to the resolution limit problem, often favoring large clusters at the expense of smaller but operationally significant groups. For fraud analytics, this creates a risk that modestly sized rings will be absorbed into broader communities.

In this study, modularity is not the primary outcome metric, but serves as a secondary check on community structure alongside the F1-score, which directly measures alignment with known fraud ring labels. This dual approach ensures that the analysis balances mathematical quality with operational relevance: modularity confirms structural coherence, while the F1-score validates whether OSINT augmentation improves the detection of actual fraud rings.

The equation for the modularity of a community is

$$Q = \frac{1}{2n} \sum_{ij} \quad \left(A_{ij} - \frac{k_i k_j}{2m}\right) \delta(c_i, c_j) \tag{1.1}$$

Where:

A_{ij} is the edge weight between nodes *i* and *j*

k_i and k_j are the sum of the weights of the edges attached to nodes *i* and *j* respectively

m is the sum of all the edge weights in the graph

c_i and c_j are the communities to which the nodes *i* and *j* belong

δ is the Kronecker delta function :

$$\delta(c_i, c_j) = \{1\ 0\ \frac{if\ c_i\ and\ c_j\ are\ the\ same\ community}{otherwise}\} \tag{1.2}$$

Partitions

Community detection algorithms work by dividing a network into partitions. In theory, the number of possible partitions grows extremely quickly with graph size, but in practice algorithms seek the partition that optimizes a quality function, such as modularity. A partition is a grouping of a network into communities such that each node belongs only to one community. In this study, partitions generated by the Leiden algorithm are evaluated against confirmed fraud ring memberships using the F1-score.

There are several community detection methods that are classified into categories based on the strategy used to identify the clusters. Bridge removal, modularity optimization, label propagation, and stochastic block modeling are the most common strategies.

Graph-based fraud detection methods

Graph-based methods have become central to fraud analytics because they capture the relationships among entities. By modeling providers, patients, and organizations as nodes

connected through claims activity, these approaches can reveal hidden structures that traditional transaction-level methods often miss. Community-based fraud detection methods are the most widely used graph-based anomaly detection method followed by probabilistic-based methods. More recent studies examined structural-based techniques and compression-based or decomposition-based approaches (Pourhabibi et al., 2020). Prior studies have largely relied on internal claims data alone. This means that relationships central to fraud schemes, such as shared business addresses, overlapping corporate ownership, or reused digital identifiers are absent from the analysis. As a result, graph-based methods risk overlooking communities that appear normal in claims data but are linked through external attributes. This gap directly motivates the present study: by enriching Medicare claims graphs with targeted OSINT indicators, it tests whether community detection performance improves when external, high-signal relationships are incorporated into the network.

This gap directly motivates the present study: by enriching Medicare claims graphs with targeted OSINT indicators, it tests whether community detection performance improves when external, high-signal relationships are incorporated into the network.

Graph Data Enrichment

A recurring limitation in graph-based fraud detection is that models often rely on a single data source, leaving important relationships unobserved. Graph data enrichment addresses this limitation by adding new nodes, edges, or attributes to an existing graph, thereby improving its capacity to represent real-world complexity. Enrichment has been used in domains such as business intelligence, cybersecurity, and recommendation systems to reveal connections that are not visible in internal datasets alone.

For fraud analytics, enrichment is especially valuable because fraudsters intentionally fragment their activity across multiple entities to avoid detection. By integrating external data, investigators can reconstruct linkages that claims records may obscure. The literature highlights several enrichment strategies, including incorporation of attribute data, bipartite structures, and heterogeneous networks, each of which increases the informational content of the graph (Leskovec et al., 2010; O'Malley et al., 2023).

However, prior research has not systematically applied enrichment to Medicare fraud networks using publicly available intelligence sources. Most existing studies enrich graphs with additional claims-based features rather than external indicators. This leaves a gap where collusive relationships that exist outside of claims data such as shared addresses, corporate ownership structures, and digital identifiers remain untested as potential community-forming attributes.

Taken together, the literature suggests that: (a) targeted external edges should improve alignment between detected communities and confirmed rings, and (b) embedding-based simplification may retain that improvement with better scalability. This motivates the study's tests of (1) F1-score gains from OSINT-augmented graphs using Leiden and (2) whether FastRP+KNN preserves comparable F1 while reducing complexity.

Homophily Network Theory

A central premise of social network analysis is homophily, the tendency of actors with similar attributes to connect with each other. In theory, this principle suggests that fraudsters should naturally cluster in provider networks: entities with shared business models, service types, or operational strategies gravitate toward one another. Fraud networks are not always assortative;

they often exhibit disassortative properties, where fraudsters connect with benign entities to camouflage their activities. This disassortativity complicates detection, as traditional methods relying on homophily assumptions may fail to capture these heterophilic interactions. In fraud networks, first-order neighbors of fraudulent entities are often heterophilic, while second-order neighbors tend to be homophilic. This structure arises as fraudsters use intermediary accounts to transfer funds, creating clusters of similar nodes at higher orders (Wang et al., 2024). Fraudsters often form communities to facilitate illegal activities. These communities may exhibit strong internal connections, making community detection algorithms effective in identifying fraudulent groups.

Homophily in Medicare Claims

Ran et al. (2024) combined social network analysis with Medicare Part D claims data to test whether there was evidence of homophily in risky prescribing among 35,000 physicians in Ohio.

Homophily in Medicare Fraud Rings

Homophily is a fundamental concept in social network analysis, referring to the propensity of individuals to form connections with others who share similar attributes, beliefs, or behaviors. In the context of Medicare fraud, homophily can manifest in various ways, such as physicians with similar prescribing patterns forming networks that may inadvertently or deliberately facilitate fraudulent activities. For instance, a study examining physician networks in Ohio found evidence of homophily in prescribing and deprescribing behaviors, suggesting that physicians with similar risky-prescribing tendencies tend to cluster together (Ran et al., 2024). The implications of homophily in Medicare fraud rings are profound. Fraudulent activities often

require coordination and trust among participants, which can be more easily established when individuals share similar characteristics or operate within the same social circles. For example, a transnational organized crime network that defrauded the U.S. Medicare system out of over $100 million was found to have a structured hierarchy with clear roles and relationships, suggesting a high degree of homophily among its members (Meyers, 2017). Moreover, homophily can influence the diffusion of fraudulent strategies within a network. A study analyzing the diffusion of potentially fraudulent Medicare home health care billing patterns found that patient-sharing across home health care agencies provided a mechanism for the rapid spread of fraudulent strategies (O'Malley et al., 2023).

Information Asymmetry Theory

While homophily theory suggests that fraudsters should naturally cluster, information asymmetry theory explains why those clusters often remain hidden. Information asymmetry occurs when one party in a system holds more or better information than another, creating opportunities for exploitation (Akerlof, 1970). In the Medicare context, providers know more about their operations, ownership structures, and billing practices than investigators or claims processors, enabling them to conceal collusion. Information asymmetry theory is a foundational concept in economics and finance that describes situations where one party in a transaction possesses more or better information than the other. This imbalance can lead to inefficiencies, fraud, and dishonesty, undermining market mechanisms and organizational performance.

This imbalance is amplified by systemic constraints. The Office of the Inspector General (OIG) faces a backlog of fraud referrals, Medicare Administrative Contractors (MACs) operate in silos, and claims must be processed within strict time windows, limiting review depth.

Together, these factors mean that many fraudulent ties such as shared addresses, overlapping corporate ownership, or coordinated digital activity—are invisible in claims-only analysis.

Prior fraud theories reinforce this dynamic. The fraud triangle and fraud diamond emphasize opportunity as a driver of misconduct, and information asymmetry creates precisely that opportunity by reducing the likelihood of detection (Isahak et al., 2023). The result is a low-certainty, high-reward environment where collusive networks can thrive.

Extending and applying Akerlof's work, Michael Spence and Joseph Stiglitz further developed the theory in the 1970s. Spence introduced the concept of signaling, where individuals or firms send signals to reduce information asymmetry, while Stiglitz explored the role of screening mechanisms to mitigate adverse selection (Hall, 2014; Klein et al., 2002). The theory gained momentum in the 1980s and 1990s as researchers applied it to various fields, including corporate finance, accounting, and behavioral economics. The Nobel Prizes awarded to Akerlof, Spence, and Stiglitz in 2001 recognized the significance of their contributions to the field of information asymmetry.

For detection methods, this means that whether fraud actors are clustered together in a network often depends less on their behavior and more on the visibility of their connections. Reducing asymmetry by introducing external intelligence sources can make hidden homophily detectable. This study operationalizes that approach by integrating three OSINT indicators: shared addresses, corporate ownership ties, and digital identifiers-- into Medicare claims graphs, testing whether they improve alignment between detected communities and confirmed fraud rings.

Fraud and Dishonesty

Information asymmetry is a key driver of fraudulent behavior in various contexts. In financial markets, it enables insider trading, where individuals with access to privileged information exploit their advantage to make profitable trades (Huddart & Ke, 2007). Similarly, in corporate settings, managers may manipulate financial statements to mislead stakeholders, a phenomenon known as earnings management (Tri Wijaya & Herwiyanti, 2023). Theoretical frameworks such as the Fraud Triangle Theory and the Fraud Diamond Theory highlight how information asymmetry creates opportunities for fraud. These theories emphasize factors like pressure, opportunity, and rationalization as key drivers of fraudulent behavior (Isahak et al., 2023).

Mitigating Information Asymmetry

To address the challenges posed by information asymmetry, researchers and practitioners have developed various strategies. These include signaling mechanisms, such as voluntary corporate disclosures, and screening mechanisms, such as third-party ratings and audits (Shishkov et al., 2022). Technological advancements, particularly in the context of Industry 4.0 (IR4.0), have also been leveraged to reduce information asymmetry. Blockchain technology, for instance, enables decentralized and transparent record-keeping, promoting faithfulness in financial reporting and enhancing market competitiveness (Yaacob et al., 2024).

Adverse Selection and Moral Hazard

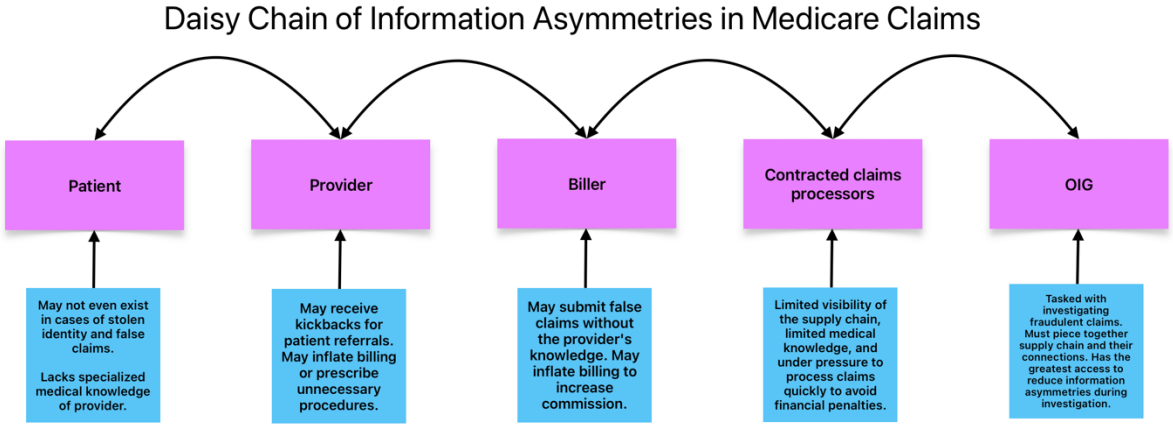
Adverse selection and moral hazard are two primary consequences of information asymmetry. Adverse selection occurs when one party is more likely to engage in a transaction because they possess superior information, leading to a disproportionate representation of less

desirable outcomes. Moral hazard, on the other hand, refers to the tendency of individuals to take greater risks when they are insulated from the consequences of their actions (Zhang, 2024).

Information Asymmetry Daisy Chains

In finance, a daisy chain refers to collusion between actors to manipulate market prices. In Medicare fraud, a daisy chain of information asymmetries creates multiple risk points for moral hazard as depicted in Figure 5. Moral hazard is a situation where an economic actor is willing to take greater risks because they do not bear the full consequences of their actions. The OIG lacks sufficient resources to investigate the volume of Medicare fraud reports it receives, and several years may pass between the fraudulent activity and a conviction. The low likelihood of getting caught and non-celerity of punishment creates a low-risk, high-profit system that is ripe for abuse.

Figure 2-7
Daisy Chain of Information Asymmetries



OSINT

Open Source Intelligence (OSINT) refers to the process of collecting and analyzing publicly available information to aid in decision-making. In the context of fraud detection,

OSINT can provide valuable insights into the relationships between actors in a network. Despite its modern applications, one of the first attempts to methodically collect OSINT was during World War Two during which the US established the Foreign Broadcast Information Service to monitor public broadcasting from unfriendly nations (Browne et al., 2024).

OSINT in Fraud

Open Source Intelligence (OSINT) has emerged as a critical tool in the detection and prevention of fraud. By leveraging publicly available data, OSINT enables organizations to gather, analyze, and correlate information from diverse sources, such as social media, news articles, and government reports. This intelligence is particularly valuable in identifying potential fraud risks, monitoring suspicious activities, and supporting investigative efforts (Chalicheemala & Chalicheemala, 2022). Financial institutions leverage OSINT to monitor transactions, assess customer risk, and detect fraudulent activities (Jesus et al., 2023). Law enforcement use OSINT to support criminal investigations, identifying patterns and tracking illegal activities (Akhgar et al., 2017). Organizations use OSINT to protect against cyber threats, monitor competitors, and safeguard brand reputation (Yadav et al., 2023).

Challenges With OSINT

One of the significant challenges is the presence of misinformation and unreliable data. Social networks and communication media are often filled with subjective opinions, fake news, and misleading information. This can complicate the OSINT process, as it is crucial to ensure that the information used is accurate and trustworthy to avoid drawing incorrect conclusions (Pastor-Galindo et al., 2020). The use of publicly available data raises ethical questions regarding privacy. There is a fine line between gathering intelligence for security purposes and infringing

on individuals' rights. Ensuring compliance with legal and ethical standards is a critical challenge for OSINT practitioners. The information landscape is constantly changing, with data being updated or removed frequently. This dynamic nature makes it challenging to maintain an accurate and up-to-date understanding of the situation being analyzed. Analysts must continuously adapt their methods to keep pace with these changes.

OSINT Integration in Graph Models

Open-Source Intelligence (OSINT) has emerged as a critical component in data enrichment for graph models, particularly in fraud detection. By leveraging publicly available data, OSINT can enhance the accuracy and comprehensiveness of graph-based fraud detection systems. Graph models, which represent entities and their relationships, benefit from OSINT by incorporating additional nodes and edges that may not be present in internal datasets. For instance, social media data, public records, and online forums can provide insights into fraudulent activities that are not immediately apparent from internal claims data

CHAPTER 3 METHODOLOGY

Introduction

This study is grounded in a postpositivist epistemology. The postpositivist view holds that empirical observations are always partial and fallible, particularly in complex systems characterized by incomplete, fragmented or intentionally obscured information such as Medicare fraud rings. This orientation directly motivates the methodological approach and the strategic use of triangulation. By combining scalable quantitative methods with qualitative intelligence such as OSINT, the study leverages multiple data sources to narrow information gaps, increase the completeness of relational signals, and provide evidence, rather than certainty, regarding the effect of augmenting Medicare claims with OSINT on community clustering within graph databases.

The purpose of this quasi-experimental quantitative study is to evaluate whether augmenting Medicare provider graphs with Open-Source Intelligence (OSINT) improves the community detection of confirmed fraud rings. By enriching provider network data with targeted external attributes such as shared addresses, corporate ownership ties, and digital identifiers – this study tests whether reducing information asymmetry leads to more accurate clustering outcomes. A deeper understanding of the relational patterns present in prior fraud ring convictions may reveal critical gaps in existing datasets and inform strategies to better link related entities within the Medicare claim supply chain.

This chapter outlines the methodological approach used to investigate the study's objectives. It begins with a restatement of the research purpose and justification for the selected quasi-experimental design. The chapter then details the data sources, graph construction process,

experimental procedures, and analytic techniques employed to evaluate the impact of OSINT augmentation. Finally, the chapter specifies the evaluation criteria. Consistent with the pyramid of evidence principle applied throughout this dissertation, only results that directly address the two research questions are emphasized: (1) whether OSINT augmentation improves fraud ring clustering, and (2) whether FastRP embeddings preserve detection accuracy while reducing graph complexity.

Research Design and Rationale

This study adopts a quasi-experimental quantitative design to examine the effect of augmenting Medicare claims graphs with Open-Source Intelligence (OSINT) on community detection performance. Specifically, the intervention consists of enriching graphs with relational data derived from publicly available sources, and measuring changes in community detection outcomes using the F1 score. By comparing baseline (non-augmented) and OSINT-augmented graph models, the study evaluates whether external data integration leads to improved detection of known fraud rings.

A true experimental design is not feasible because fraud ring memberships cannot be randomly assigned; they are determined by prior enforcement actions and exclusion lists. The quasi-experimental approach allows for systematic comparison between two naturally occurring conditions:

1. Baseline graphs, constructed solely from Medicare claims data, and;
2. OSINT-augmented graphs, enriched with shared addresses, corporate ownership ties, and digital identifiers.

The rationale for this design is twofold. First, it ensures that any differences in community detection outcomes can be attributed to the addition of OSINT features rather than to differences in graph construction or algorithm choice. Second, it provides a rigorous framework for evaluating changes in clustering accuracy using the F1-score, which measures alignment with known fraud labels.

This design is consistent with prior research in fraud analytics and network science, where quasi experimental comparisons are used to evaluate the effect of introducing a new feature or enrichment strategies (Aven, 2015; Barone & Coscia, 2018; Bauder et al., 2017; Hancock et al., 2023; O'Malley et al., 2023; Sarvari et al., 2014; Shekhar et al., 2023). By holding constant the community detection algorithm (Leiden) and varying only the informational content of the graphs, the study isolates the contribution of OSINT as an analytic intervention. This balances methodological rigor with practical feasibility, ensuring that results are interpretable, replicable, and directly responsive to the two guiding research questions.

Data Sources

This study draws on two primary categories of data:

(1) Medicare claims data, which provides the baseline for network construction, and

(2) Open-Source Intelligence (OSINT) data, which supplies targeted enrichment attributes.

The Medicare claims data used in this study includes both Part B (provider services, outpatient care, supplies) and Part D (prescription and durable medical equipment claims) for the years 2019-2021. These data provide the transactional backbone of the network, capturing provider–billing relationships that form the baseline graph structure.

The OSINT data consists of publicly available information that corresponds to three targeted indicators:

(1) Shared addresses obtained from business registrations, licensing databases, and publicly listed provider contact information,

(2) corporate ownership and registration ties derived from state-level Secretary of State filings and corporate registries, and

(3) digital identifiers, including reused business names, domains, email addresses, and phone numbers

Together, these data sources enable the construction of two comparable graph models: one derived from claims data alone and one augmented with OSINT attributes. This dual sourcing is central to the quasi-experimental design, as it isolates the effect of OSINT enrichment on community detection outcomes while holding the analytic methods constant.

Procedure

This study uses a quasi-experimental design to evaluate the effect of OSINT augmentation on the accuracy of community detection in fraud ring identification. The key analytic technique is community detection using the Leiden algorithm, a modularity-based approach well-suited for detecting densely connected subgroups in large networks.

For each identified fraud ring, two graphs were analyzed:

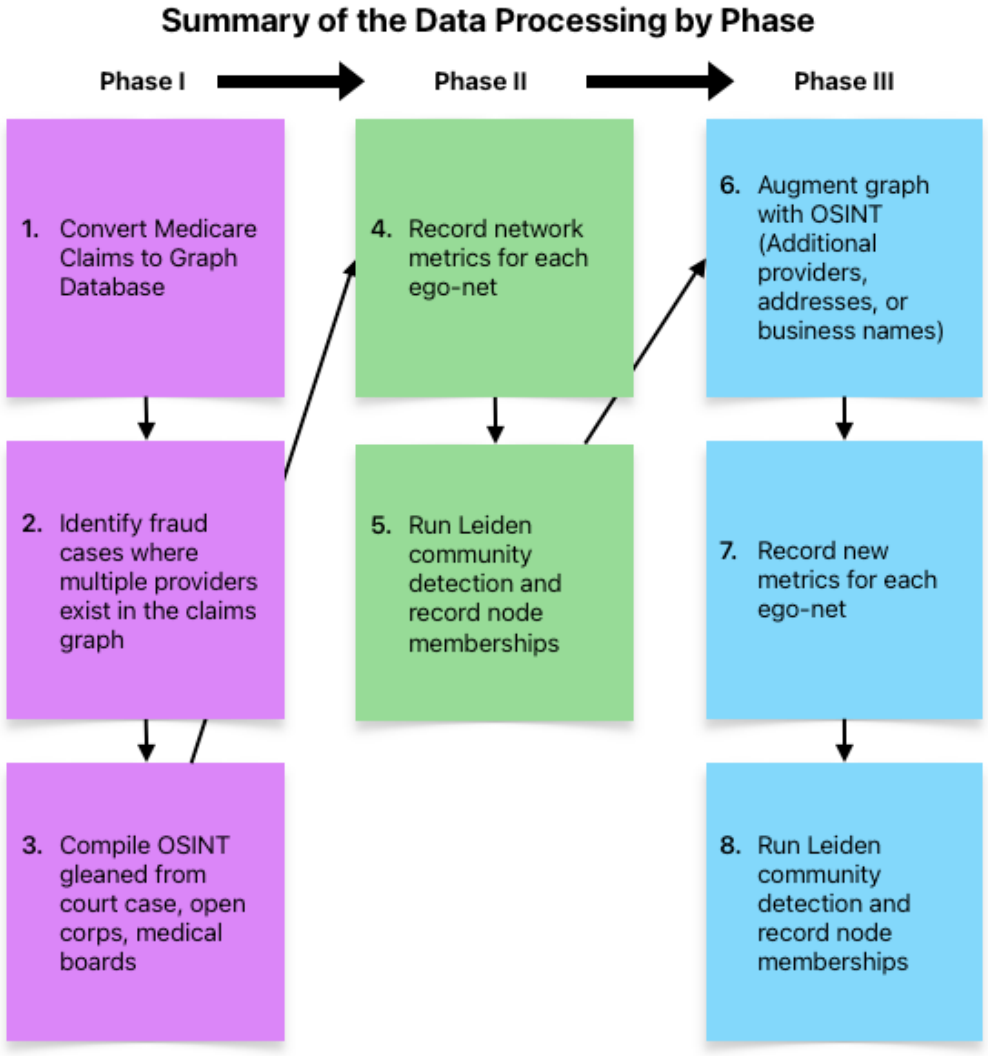
1. The baseline graph, built using claims-only data, and

2. The augmented graph, which included OSINT-derived edges.

The performance of community detection was measured using the F1 score, a harmonic mean of precision and recall, by comparing the detected community membership to the ground truth (known fraud ring members from DOJ/OIG sources). This allowed for direct comparison of how well each graph version captured the actual fraud network structure.

The data analysis followed an eight-step, three-phase process summarized in Figure 3-1 with the detailed graph construction steps in the subsequent sections. These phases reflect the pipeline from data preparation to OSINT augmentation and final performance measurement:

Figure 3-1
Summary of the Data Processing by Phase



Data Preparation and Graph Construction

The data preparation process consisted of several key steps:

1. **Convert Medicare Claims to Graph Database:** FY 2019-2021 Medicare claims data and related information including:
    a. *Part B (fee for service),*
    b. *Part D (prescribers),*

c. *Part D (durable medical equipment by provider),*

d. *Part D (durable medical equipment by supplier),*

e. *OIG List of Excluded Individuals/Entities,*

f. *NPI Provider Database*

These datasets were filtered and combined using Python and then merged into a Neo4J graph database for network analysis. Network analysis was done in both Neo4J and Gephi.

2. **Case Selection:** Fraud cases where the provider filed Medicare claims in FY2019-2021 were selected, with an emphasis on identifying cases where at least one codefendant exists in the same dataset. This criterion ensured that the subgraph similarities before and after augmentation with OSINT could be compared. The selected cases serve as representative examples of common Medicare fraud schemes, enabling a robust evaluation of OSINT's impact on network metrics and community detection.

Table 3-1
Case Selection Summary and Naming Convention

Case #	Case Name as Referenced in this Chapter	% of Nodes $n = 100$
1	First Choice Laboratory and Sonoran Desert Pathology	7%
2	J. Stanton	20%
3	Canova	8%
4	Big DME	44%
5	Stchastlivtseva	6%
6	Amity	15%

3. **Graph Augmentation:** The study utilized three targeted OSINT indicators: (1) shared addresses, (2) corporate ownership ties, and (3) digital identifiers to enhance the initial claims graphs. OSINT data corresponding to these indicators were gathered from multiple publicly available sources, including:
    a. Court documents
    b. Medical license board information
    c. Google Reviews
    d. Business registry databases

    OSINT-derived edges and weights include:

Table 3-2
OSINT-Edge Weightings

Edge	Weight
[:OWNS]	10,000
[:BUSINESS_ASSOCIATE]	10,000
[:HAS_ADDRESS]	10,000

4. **OSINT Validation and Quality Control:** To ensure data reliability, each OSINT indicator was verified through at least two independent public sources, including court filings, before inclusion. Shared addresses were cross-checked between Secretary of State filings and NPI Registry entries; corporate ownership ties were confirmed using state business filings and officer or agent names; and digital identifiers (phone numbers, domains, business names) were validated via WHOIS, web archives, and business directory cross-matches. Each OSINT-derived edge was created only when the relationship could be explicitly verified

through public documentation (e.g. shared address, verified business partnership, or registered ownership record) rather than inferred by similarity thresholds.

5. **Quantitative Analysis:** This study compared the similarity of subgraphs (egonets) before and after augmentation with OSINT. Community detection with the Leiden algorithm was conducted before and after augmentation to test whether the OSINT altered the community membership of egonets. The metrics used to compare the egonets include:
    a. Jaccard similarity
        i. Jaccard similarity is calculated as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.1}$$

where A and B are the set of elements connected to each node. The Jaccard similarity between two nodes measures the shared connections or similarity in billing behavior. This metric was used to measure overlap between nodes before and after OSINT augmentation, providing insights into how additional edges strengthen network cohesion. Jaccard similarity is a commonly used metric to quantify the similarity of two sets of elements (Travieso et al., 2024).

    b. K-nearest neighbors similarity score
        i. KNN classification using cosine similarity

$$cos\ \theta\ = \frac{A \cdot B}{\|A\|\|B\|} \tag{3.2}$$

measures angular alignment between node embeddings. KNN is an ideal clustering algorithm for learning low-dimensional manifolds (Kramer, 2011). In this study, incorporating KNN-derived edges allows for the integration of individual node characteristics, extending beyond reliance on edge weights alone.

c. Modularity (of the Leiden communities)
    i. Modularity measures the strength of a community structure by evaluating the density of connections within communities compared to the expected density if connections were random. Networks with high modularity have dense connections within communities and sparse connections between them. The Leiden community detection algorithm is a modularity-optimizing algorithm. Leiden is an ideal community detection algorithm for weighted, heterogeneous graphs (Hairol Anuar et al., 2024). Modularity is a commonly used metric in community detection used to evaluate community structure (Hairol Anuar et al., 2024).
d. F1 Score (of the Leiden communities)
    i. The F1 score is the harmonic mean between precision and recall. In this study, I consider nodes associated with fraudulent activity as ground truth nodes. Because it is not possible to know which nodes are false positives, only the community membership of ground truth nodes is considered. Therefore, Precision = 1 in the

F1 score calculation. An F1 score above .9 may indicate overfitting to the data. Scores closer to .7 are an ideal balance between overfitting and detecting granular structures.

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP)+False\ Positives\ (FP)}$$

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

$$F1\ Score = \frac{2\cdot(Precision\cdot Recall)}{Precision+Recall} \quad (3.3)$$

The F1 score is a commonly used metric in machine learning applications. In the context of community detection, it is most widely used for assessing the classification of labels in community detection (Liu et al., 2018).

e. Average and Weighted Degree

i. The mean number of connections per node

$$\underline{k} = \frac{2e}{n} \quad (3.4)$$

quantifies participants' systemic embeddedness. This metric contextualizes individual influence capacity, as nodes with higher-than-average degrees often occupy central positions. However, it should be interpreted alongside density, as degree alone does not capture edge distribution patterns (e.g., hubs vs. uniform dispersion) (Crossley et al., 2015).

ii. Average weighted degree,

$$\underline{k_w} = \frac{\sum \quad w_{ij}}{n}, \tag{3.5}$$

using dollar amounts or predetermined scores as edge weights, quantifies the economic scale of fraudulent interactions. In weighted graphs, this metric is appropriate for measuring connectivity and influence of nodes in egonets (Crossley et al., 2015).

f. Density

   i. Density measures the completeness of connections between co-conspirators within a fraud ring's egonet, calculated as

$$\frac{2e}{(n*(n-1))} \tag{3.6}$$

   In the context of this study, lower network density within networks may reflect heterophilic (dissimilar) relationships and is reported as a descriptive statistic to capture structural diversity in local neighborhoods (Crossley et al., 2015).

g. Diameter

   i. The maximal path length between any two nodes identifies operational vulnerability points. As an egonet structural metric, diameter is included as a descriptive indicator of egonet topology (Crossley et al., 2015).

h. Clustering Coefficient

   i. This metric, computed as

$$C_i = \frac{|\{e_{jk}: v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)} \quad (3.7)$$

in directed graphs, is another descriptive metric characterizing the degree of local cohesion in each egonet (Crossley et al., 2015).

Figure 3-2 depicts a simpler summarization of the data preparation and analysis process.

Figure 3-2
Summarization of Analysis Recap

Node Similarity

In this study, node similarity within the graph was computed using the K-Nearest Neighbors (KNN) algorithm implemented in Neo4j's Graph Data Science library, leveraging Fast Random Projection (FastRP) embeddings to represent nodes as vectors in a lower-dimensional embedding space. Pairwise node similarity was subsequently measured using the cosine similarity metric, enabling the identification of structurally and semantically similar nodes based on their embedding proximity. FastRP generates low-dimensional vector representations of nodes by iteratively diffusing structural features through sparse matrix factorization. Applied to fraud egonets, these embeddings encode relational patterns (e.g., provider centrality) into a latent space where similarity reflects functional equivalence. I selected dimensionality of 64 to balance computational requirements with embedding usefulness. FastRP is up to 4,000 times faster than other popular embedding algorithms such as Node2Vec (Traag et al., 2019).

Naming Convention of Measurement Stages

Community detection was measured at four stages across 43 resolutions, both weighted and unweighted (Table 3-2). Egonet analysis is included for each of the six cases at each of the four stages. The naming convention used for the stages are:

1. **Base**: non-augmented claims
2. **BaseKNN:** non-augmented claims with edges derived from KNN similarity
3. **OSINT:** claims graph augmented with OSINT
4. **OSINT_KNN:** claims graph augmented with OSINT and KNN similarity edges

Table 3-2
Measurement Stages

Stage	Resolutions	Egonets	Nodes	Total Tests
Base	43	6	100	258
BaseKNN	43	6	100	258
OSINT	43	6	100	258
OSINT_KNN	43	6	100	258
uwBase	43	6	100	258
uwBaseKNN	43	6	100	258
uwOSINT	43	6	100	258
uwOSINTKNN	43	6	100	258
Combined Total	**344**	**48**	**800**	**2,064**

Unit of Analysis Clarification

Clarification of the unit of analysis is outlined in Table 3-3.

Table 3-3
Clarification of Unit Analysis

Term	Statistical Role	Definition
Nodes (N=100)	Observational units	Individual providers/suppliers
Cases/Egonets (N=6)	Primary sampling units	Fraud networks and 1st degree connections
Trials (N=2,064)	Sensitivity tests	Leiden community detection at incremental resolutions

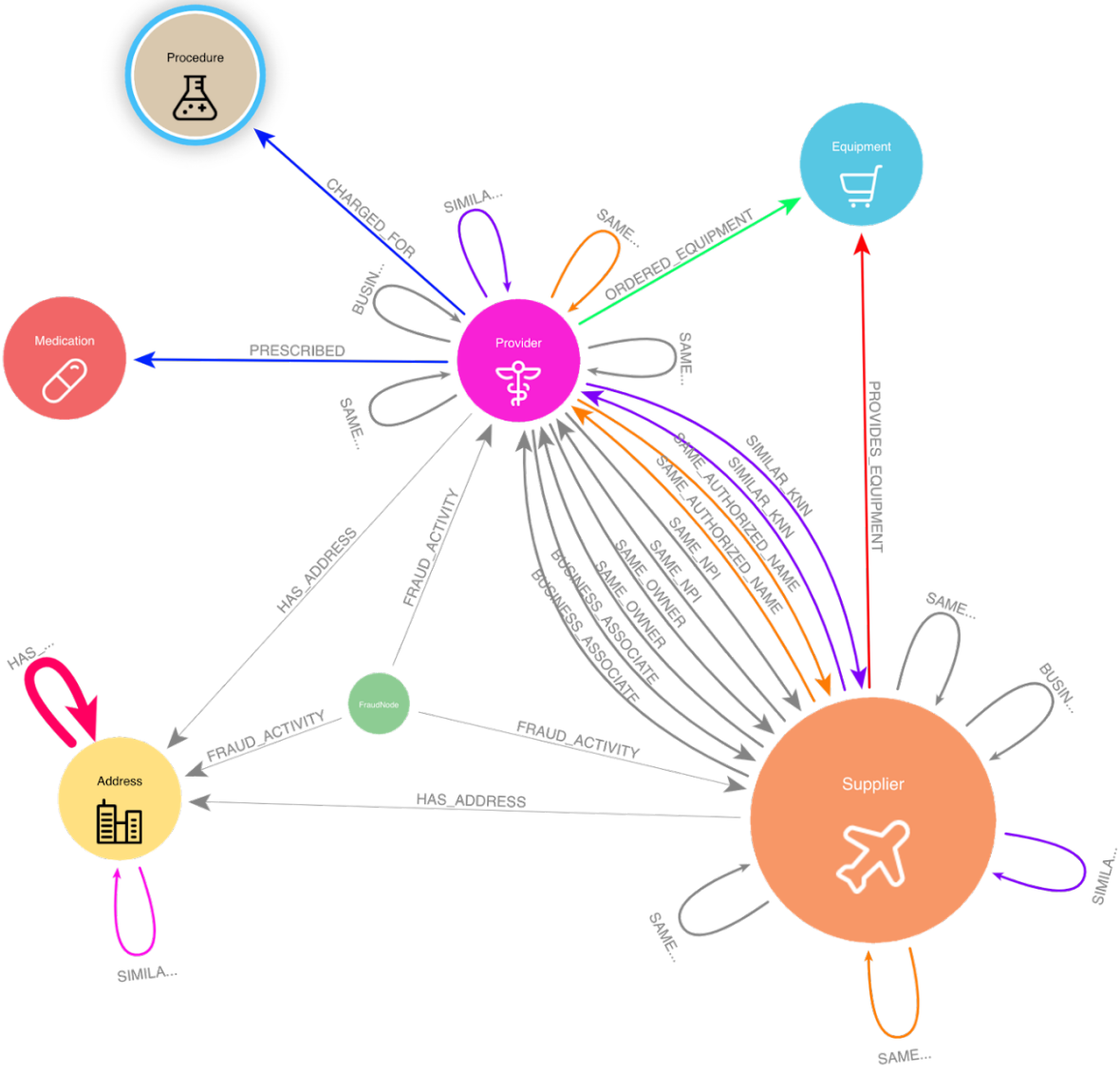
Graph Topology (schema)

Figure 3-3 shows the graph database topology [or schema]. The primary keys for each node type are in Table 3-4.

Table 3-4
Primary Keys

Node Type	Primary Key	Notes
Provider	Name+NPI	Name concatenated with NPI
Supplier	Name+NPI	Name concatenated with NPI
Procedure	Hcpcs_cd	Code assigned by CMS
Equipment	Hcpcs_cd	Code assigned by CMS
Medication	Generic_Name	Generic name of the medication
Address	Address	Due to variations in addresses, a second addressParts field was created along with a SIMILAR_ADDRESS relationship for nodes with the same addressParts and zip code
FraudNode	NPI	These nodes were matched to NPI numbers or matching addresses from the exclusion list

Figure 3-3
Graph Database Schema Used



Edge Weightings

Connections between providers or suppliers are not all equally significant. Two users with Facebook accounts would not be a significant linked attribute. A shared phone number or address between two users is a much more significant connection. Similarly, two providers who prescribe the same medication is of much lower significance than two providers sharing an address. The exception to this is if few providers bill for a specific service, particularly if providers are in the top quartile of billers for that service. Unlike unweighted approaches, weighted graphs allow Leiden to differentiate between strong and weak ties, emphasizing edges that reflect more significant or frequent relationships. This leads to communities that are not only structurally cohesive but also semantically meaningful, particularly in networks where edge weight conveys critical context—such as transaction volume, communication frequency, or similarity scores (Traag et al., 2019). Using this logic, the graph edges were weighted as outlined in Table 3-5.

Table 3-5
Edge Weightings

Relationship	Weight (composite score)
[:HAS_ADDRESS]-()-[:HAS_FRAUD]	Exclusion weight * 10,000
[:HAS_FRAUD]	Exclusion weight * 25,000
[:HAS_ADDRESS]	1,000
[:PRESCRIBED]	Cost
[:PRESCRIBED] (controlled substance)	Cost * 2.5
[:ORDERED_EQUIPMENT]	Cost
[:SUPPLIES_EQUIPMENT]	Cost
[:SIMILAR_ADDRESS]	2,500
[:SAME_NUMBER]	100,000
[:SIMILAR_KNN] (Cosine)	Score * 1,000
[:SAME_AUTHORIZED_NAME]	1,000
[:SAME_NPI]	100,000

LEIE Exclusions Weights

The List of Excluded Individuals and Entities database provides the name, NPI, address, exclusion reason, and other relevant details about excluded providers. Exclusion weights were assigned as outlined in table 3-6.

Table 3-6
Exclusion Codes and Reasons

Exclusion Code	Exclusion Reason	Exclusion Weight
1128a4	Felony conviction relating to controlled substance. Minimum Period: 5 years	8
1128a3	Felony conviction relating to health care fraud. Minimum Period: 5 years	10
1128a1	Conviction of program-related crimes. Minimum Period: 5 years	10
1128b4	License revocation, suspension, or surrender. Minimum Period: Period imposed by the state licensing authority.	8
1128b7	Fraud, kickbacks, and other prohibited activities. Minimum Period: None	10
1128b6	Claims for excessive charges, unnecessary services or services which fail to meet professionally recognized standards of health care, or failure of an HMO to furnish medically necessary services. Minimum Period: 1 year	6
1128a2	Conviction relating to patient abuse or neglect. Minimum Period: 5 years	4
1128b1	Misdemeanor conviction relating to health care fraud. Baseline Period: 3 years	3
1128b14	Default on health education loan or scholarship obligations. Minimum Period: Until default or obligation has been resolved.	1
1128b8	Entities controlled by a sanctioned individual. Minimum Period: Same as length of individual's exclusion.	5
1128b5	Exclusion or suspension under federal or state health care program. Minimum Period: No less than the period imposed by federal or state health care program.	4

Case Selection Rationale

The selection of six cases for this study was guided by two primary considerations: diversity in fraud typologies and methodological feasibility. The chosen cases represent common Medicare fraud mechanisms—such as kickbacks, upcoding, unnecessary services, durable medical equipment (DME) schemes, and identity theft—ensuring that findings are applicable across multiple scenarios. This approach allows for an evaluation of OSINT's impact on graph-

based fraud detection across heterogeneous network structures. While a larger sample size could enhance generalizability, the time-intensive nature of OSINT data collection and graph augmentation necessitated a focused analysis prioritizing analytical depth over breadth. Similar studies in Medicare fraud detection have demonstrated that small but representative samples can yield meaningful insights (Bauder & Khoshgoftaar, 2018).

Medicare fraud cases face inherent data challenges: exclusion lists fail to link co-defendants due to staggered legal proceedings (e.g., some defendants are flagged post-conviction, while others remain unflagged pending appeals) and in some cases, business entities are dissolved rather than excluded. To mitigate this fragmentation, cases were selected using two criteria:

1. Presence of at least one co-defendant in claims dataset, ensuring baseline connectivity for pre/post-OSINT comparisons.
2. Diverse fraud typologies reflecting common schemes identified in DOJ prosecutions (GAO, 2024):

Table 3-7
Medicare Fraud Typologies Matrix

Scheme	Case 1	Case 2	Case 3	Case 4	Case 5
False Claims	√	√		√	√
Upcoding	√	√	√		
Unnecessary Services	√	√	√	√	√
Kickbacks	√	√	√		
DME	√			√	√
Identity Theft				√	√

This approach tests OSINT's efficacy across heterogeneous fraud mechanisms while controlling for shared network attributes (e.g., co-defendant edges). By selecting cases where traditional

claims data omits critical actors—particularly non-billing intermediaries—the methodology mirrors real-world detection challenges outlined in Chapter 1.Population and Sampling

The sample population for this study consisted of Medicare providers who submitted claims under Medicare Part B (fee-for-service) or Medicare Part D (prescriptions and durable medical equipment). Due to computational constraints, only claims from the 2019–2021 datasets were included in the analysis. Since most providers are not confirmed fraud actors, a sampling strategy was required to identify cases suitable for evaluating the effect of OSINT augmentation on community detection.

This study employed criterion-based sampling. Providers were included in the evaluation set if they met the criterion of confirmed fraud involvement, as documented through official exclusion from Medicare or enforcement actions reported by the Department of Justice (DOJ) and the Office of Inspector General (OIG). These confirmed cases served as the ground truth labels against which community detection outcomes were validated.

Providers were identified by their National Provider Identifier (NPI), which can be assigned to either individuals or organizations. Fraudulent actors were identified using the Medicare List of Excluded Individuals/Entities (LEIE), which contains NPIs for all providers currently excluded from the Medicare program. As of February 10, 2025, the LEIE included 81,472 excluded providers. Only those NPIs—and associated addresses—that appeared in the 2019–2021 claims datasets were retained for analysis.

To enrich the dataset, the NPI Lookup database was used to supplement address information and identify additional authorized signers for organizational providers. Data cleaning and preprocessing were conducted in Python. Given that the Medicare claims data are

pre-aggregated before public release, missing values were not expected. The only anticipated exception was the firstName field for organizational NPIs, which do not require an individual name. This assumption was confirmed programmatically. For consistency, organizational records with an empty firstName field were filled with the placeholder value "biz" during import into Neo4j.

To ensure case-insensitive matching, all names and addresses were converted to lowercase during import. Address standardization presented greater challenges due to formatting variations. To mitigate this, an addressParts field was created by extracting the first three to four segments of each space-delimited address (e.g., “p o box 123,” “123 n main,” “123 north main”). A custom SIMILAR_ADDRESS relationship was then generated in the graph database to connect potentially duplicated or closely matching addresses. Next, the files were imported to Neo4J and mapped according to the schema in Figure 3-3.

Unique constraints were created for each node and relationship to prevent duplicates. Details from the claim data such as cost were added to the relationship values, and aggregate totals for providers were added as properties in nodes. The node properties do not influence community detection, but the cost values were used as weights.

Data Understanding

The Centers for Medicare and Medicaid Services (CMS) provides free API access to download claims data by year. CMS also provides accompanying data dictionaries and methodology documentation explaining how the data is sourced and compiled. Data year reflects claims from July 1 of the year listed through June 30 of the subsequent year. A dataset listed as 2021 includes claims received by Medicare between July 1, 2021 and June 30, 2022. Table 3-8

outlines the proposed baseline datasets. To keep the file sizes manageable without the need for cloud computing services, I will combine at most 3 years of records (2019-2021).

The List of Excluded Individuals/Entities (LEIE) provided by the Office of the Inspector General (OIG) contains the list of registered Medicare providers who are currently excluded from Federally funded healthcare programs for Medicare fraud or convictions related to fraud in other healthcare programs, unlawful distribution or dispensing of controlled substances, or medical license revocation/suspension tied to professional misconduct such as receiving unlawful kickbacks, controlling a sanctioned entity as an owner, officer, or managing employee. These individuals will serve as our fraud nodes/labels in the baseline model. We cannot assume that the list of excluded entities is exhaustive as this would be like assuming that everyone who has ever committed a crime has been caught and convicted. The row and column counts listed in Table 1 are the raw-unprocessed dataset values retrieved directly from CMS.

Table 3-8
Datasets Used

Data Set Name	Description	Year	Version	Rows	Columns
Medicare Part B Outpatient	Physician & Other Practitioners - by Provider and Service	2019	867b8ac7-ccb7-4cc9-873d-b24340d89e32	10,140,228	28
Medicare Part B fee-for-service	Physician & Other Practitioners - by Provider and Service	2020	c957b49e-1323-49e7-8678-c09da387551d	9,449,361	28
Medicare Part B fee-for-service	Physician & Other Practitioners - by Provider and Service	2021	31dc2c47-f297-4948-bfb4-075e1bec3a02	9,886,177	28
Medicare Part B - DMEDS	Medicare Durable Medical Equipment, Devices & Supplies - by Referring Provider and Service	2019	eb0019f6-791d-4065-ae4e-4761d2f6c9f2	1,656,449	33

Data Set Name	Description	Year	Version	Rows	Columns
Medicare Part B - DMEDS	Medicare Durable Medical Equipment, Devices & Supplies - by Referring Provider and Service	2020	323df359-ceac-4525-a350-e2cd9eb128fe	1,607,435	33
Medicare Part B - DMEDS	Medicare Durable Medical Equipment, Devices & Supplies - by Referring Provider and Service	2021	46ae675c-bc81-40ca-aa79-64da1c1ec9d9	1,516,153	33
Medicare Part D Prescribers	Part D Prescribers by provider and Drug	2019	2a6705e6-7a1e-460c-ba22-35249a531918	25,401,870	22
Medicare Part D Prescribers	Part D Prescribers by provider and Drug	2020	7795fe20-e80e-435a-a9ed-d2d65e05feeb	25,209,729	22
Medicare Part D Prescribers	Part D Prescribers by provider and Drug	2021	ab29d858-269a-4d97-908f-a26b1cf95f61	25,231,862	22
Medicare Part B DME	Medicare Durable Medical Equipment, Devices & Supplies - by Supplier and Service	2019	5165dbdd-eae3-4b94-82cd-da635e00e0fa	575,223	32
Medicare Part B DME	Medicare Durable Medical Equipment, Devices & Supplies - by Supplier and Service	2020	1fe3a8c3-ecfd-44b9-9418-fda1843f72a5	542,090	32
Medicare Part B DME	Medicare Durable Medical Equipment, Devices & Supplies - by Supplier and Service	2021	e531ffc9-57ad-48db-8d88-f686edb8b8e1	508,052	32
LEIE Database	https://oig.hhs.gov/exclusions/exclusions_list.asp	2025	February 2025	Only matched NPIs	
NPI Database	https://download.cms.gov/nppes/NPI_Files.html	2025	NPPES	Only addresses for matched NPIs	

Data Analysis and Tools

The analysis proceeded in three steps: community detection, evaluation, and scalability testing. Each step was designed to isolate the effect of OSINT augmentation while directly addressing the two research questions.

Both baseline and OSINT-augmented graphs were partitioned using the Leiden algorithm, selected for its scalability, stability, and ability to produce well-connected communities (Traag et al., 2019). Resolution parameters were tuned iteratively to test sensitivity to community size, with adjustments recorded for transparency. The primary evaluation metric was the F1-score, which measures the alignment between detected communities and confirmed fraud ring memberships. This metric was chosen because it balances precision (avoiding false positives) and recall (capturing true positives), ensuring that results are operationally relevant. Modularity was also recorded as a secondary measure of partition quality, but interpretive weight was placed on the F1-score to maintain focus on the study's central objective.

To address the second research question, OSINT-augmented graphs were embedded using the FastRP algorithm. A K-Nearest Neighbor (KNN) graph was then reconstructed from cosine similarity in the embedding space. This allowed comparison of community detection results between full OSINT-augmented graphs and their embedded counterparts. The key analytic test was whether embeddings preserved comparable F1-score performance while reducing graph complexity.

These procedures ensured that each analysis step was explicitly tied to the research design: holding constant the community detection algorithm while varying graph enrichment and embedding conditions. By combining methodological transparency with targeted evaluation metrics, the study provides a rigorous test of whether OSINT reduces information asymmetry and improves fraud ring detection.

A combination of statistical and graph analysis tools was used to analyze the data. Python served as the primary tool for data cleaning, transformation, and visualization. Libraries such as Pandas and NumPy were used to manipulate tabular data, while Plotly, Gephi, and Neo4j Bloom supported visual exploration. Jamovi, an open-source statistical platform, was used to run descriptive statistics and basic hypothesis tests as needed. Graph-based analyses were performed using Neo4j, a graph database used to structure and query provider relationships, and Gephi, which was used for network visualization and centrality metric calculations. Together, these tools facilitated both quantitative analysis and the exploration of structural patterns within the fraud networks.

Data Analysis

Because no prior research provided an empirically justified estimate of the expected OSINT effect, an a priori power analysis was not feasible. Instead, a post-hoc power analysis was conducted using the observed effect size from the Wilcoxon signed-rank test, calculated as a rank-biserial correlation. The analysis incorporated 43-paired F1-scores (one per resolution) with a= .05 (two-tailed) and the observed effect size (r = 1.00). The resulting power confirmed that

the study was full powered to detect the observed difference between baseline and OSINT-augmented graphs.

To corroborate these results, a Monte Carlo simulation was conducted in Python to assess the sensitivity of the design to smaller sample sizes. Using the same effect-size distribution and Wilcoxon test structure across 1,000 simulated iterations, showed that even with just n=6 paired cases, the study would achieve full power (1-$\beta = 1.00\ ;\ a = .05$). These results demonstrate that the large observed effect size compensates for the modest number of empirical cases, confirming that the quasi-experimental design possesses sufficient sensitivity to detect meaningful effects.

The total number of paired F1-score comparisons thus provided a robust empirical basis for estimating achieved power. Summary statistics were subsequently visualized in Jamovi. Consistent with recommendations by Quach et al. (2022), the post-hoc approach was appropriate because the true effect size could only be established after empirical testing (Quach et al., 2022). Both computational methods, G*Power and Monte Carlo simulations, converged on a statistical power of 1.00 indicating that the Wilcoxon text was fully powered to detect the magnitude of the change observed between pre and post-augmentation F1 scores.

This high level of power, combined with exceptionally large effect sizes (rank-biserial correlation = 1.000 for weighted graphs, 0.980 for unweighted graphs), provides strong statistical support for the robustness of the findings despite the modest sample size. The G*Power results are included here and the code used to run the simulations is in the Appendix.

Table 3-9
Power Analysis G*Power Output

Input	**Tail(s)**	**2**
	Parent Distribution	Normal
	Effect size dz	1

	α err prob	0.05
	Total sample size	43
Output	**Noncentrality parameter**	**6.41**
	Critical t	2.02
	Df	40.10
	Power (1-β)	1.00

F1 Score and Effective F1 Score

The F1 score is the harmonic mean between precision and recall. In this study, I consider nodes associated with fraudulent activity as ground truth nodes. It is not possible to know which nodes are false positives so only the community membership of ground truth nodes is considered. Therefore, Precision = 1 in our F1 score calculation is equation 3.

High F1 scores at low resolutions can be misleading in cases where all nodes are grouped into a single community. To provide additional context for the F1 scores presented herein, I include a heuristic metric, the effective F1 score calculated as

$$F1_{effective} = F1_{score} * log\ (communityCount) \quad (8)$$

The effective score is a heuristic that aides in distinguishing artificially inflated F1 scores.

Reliability and Validity

Ensuring reliability and validity is critical for evaluating whether OSINT augmentation improves fraud ring detection. This study incorporates multiple safeguards to enhance methodological rigor and to address potential threats to accuracy, consistency, and interpretability. Reliability was strengthened through a consistent graph construction pipeline. The same claims data served as the foundation for both baseline and OSINT-augmented graphs, with enrichment applied systematically to ensure comparability. Community detection was

conducted using the Leiden algorithm across multiple runs, with parameter settings documented to ensure repeatability. The use of publicly available OSINT sources also enhances reproducibility, as enrichment features can be independently verified.

The quasi-experimental design isolates OSINT as the sole differentiating factor between baseline and augmented graphs. By holding constant the claims data, analytic algorithms, and evaluation metrics, the study minimizes confounding influences. Criterion-based sampling further enhances validity by using confirmed fraud cases as ground truth labels rather than inferred or proxy measures. While the study focuses on Medicare claims, its design principles are transferable to other large-scale fraud detection domains where collusive behavior is difficult to detect in siloed and tabular datasets. The use of OSINT indicators that are broadly accessible addresses, corporate registrations, digital identifiers, supports generalizability to other healthcare systems and financial crime contexts.

Moreover, as of this writing, some cases from 2018 and 2019 are still undergoing prosecution. This further complicates the task of identifying fraud rings whose activities occurred within the sampling window (2019–2021), since some actors may not yet appear in formal records. In many cases, indictments offered a wealth of supplemental information, revealing provider behaviors and network structures that would otherwise remain obscured.

However, the inherently varied and complex nature of fraud schemes means that even detailed open-source intelligence may fail to capture all relevant relationships between providers. Given the time-intensive nature of manual data collection, validation, and cross-referencing, this study focuses on six well-documented cases (comprising 100 nodes) selected for their investigative clarity and relevance to the study period.

The F1-score was chosen as the primary evaluation metric because it directly captures alignment with known fraud ring memberships, ensuring that the construct of "improved detection" is measured in an operationally meaningful way. Modularity was included as a secondary quality measure to confirm that community structures were coherent, but interpretive weight was placed on fraud-label alignment to preserve construct fidelity.

Together, these safeguards ensure that the study's findings are methodologically sound, reproducible, and relevant both to Medicare fraud detection and to the broader field of graph-based analytics.

Ethical Considerations

It is important to emphasize that business associations identified in the network do not imply guilt. Providers named explicitly in court records are noted in the case details; however, legal proceedings often span several years, and outcomes can vary significantly. In several reviewed cases, defendants negotiated plea deals, became informants, or had their records sealed. In a few instances, not all individuals named in the indictment had been formally charged at the time of analysis.

Data Process & Security

All data used in this study were obtained from publicly available sources. To support reproducibility and future research, four database configurations—Base, BaseKNN, OSINT, and OSINT_KNN—have been made available on GitHub. Provider names were intentionally not anonymized to allow readers to independently verify the findings and more easily construct mental models of how potential fraud rings may manifest within the network.

CHAPTER 4. RESULTS: PRESENTATION AND ANALYSIS OF THE DATA

Overview of the Study

This quasi-experimental, multilevel study examines the effect of Open-Source Intelligence (OSINT) augmentation in enhancing fraud ring detection within Medicare claims data (2019-2021). This study does not seek to generalize outcomes to all Medicare claims but instead tests the mechanical effect of graph augmentation using OSINT in well-documented fraud cases by leveraging ground-truth actors. Using a criterion-based case selection approach, this research analyzes six fraud networks comprising 100 nodes across 43 resolution parameters and four graph states, yielding 2,064 analytical trials. This methodological design aligns with graph-based fraud analytics research (Pourhabibi et al., 2020). While leveraging purposive sampling techniques to ensure inclusion of cases with characteristics essential to fraud detection evaluation. By employing non-parametric Wilcoxon signed-rank tests for hypothesis testing—a methodological choice supported by normality assessments—this study evaluates community detection metrics before and after OSINT integration. This chapter reiterates the research questions, details the analytical framework, presents in-depth egonet analyses, and synthesizes cumulative findings across all case networks.

Research Questions and Hypotheses

The study was guided by the research questions:

RQ1. Can augmenting graphs with OSINT improve the community detection of fraud rings over non-augmented graphs as measured by the F1-score?

H0. *The F1 score will not change with the addition of OSINT-derived edges.*

H1. *The F1 score will increase with the addition of OSINT-derived edges.*

RQ2. Can we preserve the same level of accuracy by applying reductionism to the OSINT-augmented graphs using fastRP embeddings?

H0. *The F1 score will not change after incorporating OSINT-derived edges and fastRP embeddings in Leiden community detection.*

H1*. The F1 score will increase after incorporating OSINT-derived edges and fastRP embeddings in Leiden community detection.*

The hypotheses were tested using the F1 score derived from the community membership results of the Leiden community detection algorithm. RQ2 Edges were generated from the cosine similarity of FastRP embeddings (64 dimensions) within the k-nearest neighbors algorithm.

RQ1. Can augmenting graphs with OSINT improve community detection of fraud rings over non-augmented graphs?

The results strongly support the alternative hypothesis that augmenting graphs with OSINT improves fraud ring detection over non-augmented graphs, leading to the rejection of the null hypothesis. Statistically significant improvements were observed across all metrics when graphs were augmented with OSINT-derived edges. Leiden clustering was conducted on weighted and unweighted graphs across all graph states, and Wilcoxon signed-rank tests (Table 4-1) revealed consistent differences:

Weighted Graphs: OSINT augmentation significantly improved F1 scores *($W=946$, $p<.001$)* with a mean increase of 0.211, representing a 21.1% improvement in classification accuracy. The effect size (rank-biserial correlation=1.000) indicates exceptionally strong directional consistency favoring OSINT augmentation across all paired comparisons.

Unweighted Graphs: Similar improvements were observed in unweighted graphs, with OSINT augmentation significantly improving F1 scores *(W=894, p<.001)* with a mean increase of 19.9% and a large effect size (rank-biserial correlation=0.980).

Effective F1 Scores: When adjusting for community count using my effective F1 score metric, the improvements were confirmed. OSINT augmentation increased effective F1 scores by 0.939 in weighted graphs and 0.779 in unweighted graphs (both p<.001). Notably, these scores retain their original scale rather than being rescaled to 0-1. Effective F1 scores were calculated to account for community count variability, penalizing oversimplified structures with few communities while preserving accuracy in larger networks. This adjustment ensures that high F1 scores reflect meaningful granularity rather than inflated values from sparse community detection.

Fig. 4-3 illustrates the relationship between F1 score and modularity across resolution parameters in weighted Leiden clustering runs. The red line (OSINT-augmented graphs) consistently outperforms the blue line (base graphs), maintaining higher F1 scores even at higher resolutions where modularity typically declines. Fig. 4-1 and 4-2 show the Leiden results side-by-side with the effective F1 scores confirming that the significant improvement in F1 scores are maintained when adjusting for community count.

While OSINT augmentation significantly improves community detection metrics across all graph states and configurations, it is important to note that these results reflect specific fraud typologies analyzed in this study. Future research should explore how OSINT impacts detection accuracy in more diverse or less structured fraud networks.

Table 4-1
Research Question 1-Statistics

Base vs OSINT Weighted, Unweighted, and Effective F1 Wilcoxon Signed Rank Test

Effective F1 = F1score * log(communityCount)

uw = Unweighted

Paired Samples T-Test

			Statistic	p	Mean difference	SE difference		Effect Size
uw_OSINT_Mean_F1	uw_Base_Mean_F1	Wilcoxon W	894 [a]	<.001	0.19896	0.0164	Rank biserial correlation	0.9801
uw_OSINT_Effective_Mean_F1	uw_Base_Effective_Mean_F1	Wilcoxon W	940	<.001	0.77865	0.0644	Rank biserial correlation	0.9873
OSINT_Mean_F1	Base_Mean_F1	Wilcoxon W	946	<.001	0.21114	0.0162	Rank biserial correlation	1.0000
OSINT_Effective_Mean_F1	Base_Effective_Mean_F1	Wilcoxon W	946	<.001	0.93948	0.0724	Rank biserial correlation	1.0000
Base_Mean_F1	uw_Base_Mean_F1	Wilcoxon W	273	0.016	−0.03366	0.0121	Rank biserial correlation	−0.4228
Base_Effective_Mean_F1	uw_Base_Effective_Mean_F1	Wilcoxon W	621	0.075	0.14033	0.0604	Rank biserial correlation	0.3129
OSINT_Mean_F1	uw_OSINT_Mean_F1	Wilcoxon W	495	0.795	0.00238	0.0119	Rank biserial correlation	0.0465
OSINT_Effective_Mean_F1	uw_OSINT_Effective_Mean_F1	Wilcoxon W	826	<.001	0.36404	0.0630	Rank biserial correlation	0.7463

Note. $H_a\ \mu_{Measure\ 1\ -\ Measure\ 2} \neq 0$

[a] 1 pair(s) of values were tied

Figure 4-1
Weighted Leiden Detection F1 Scores

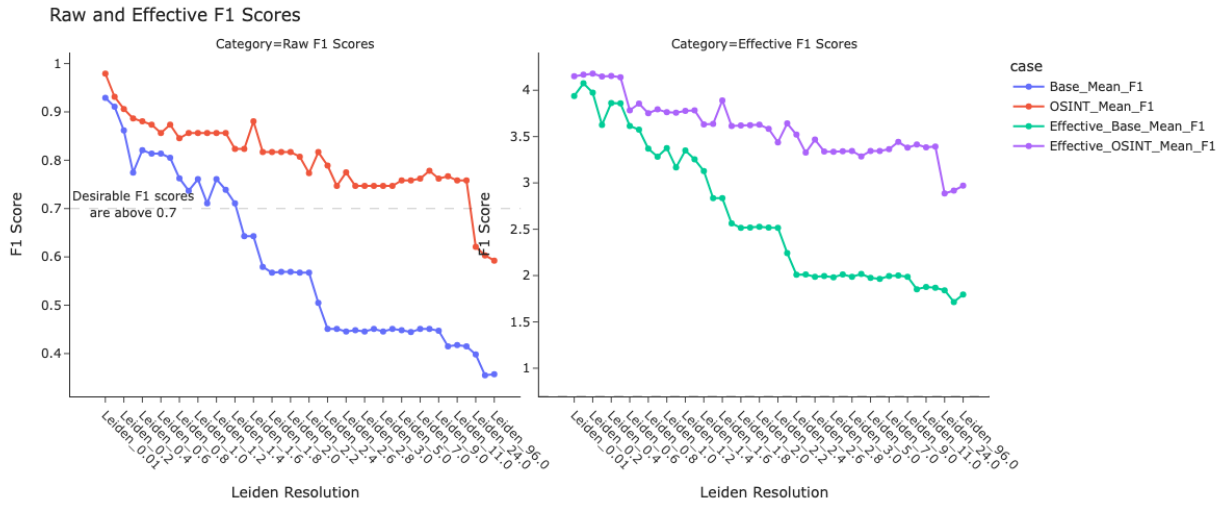


Figure 4-2
Unweighted Leiden Detection F1 Scores

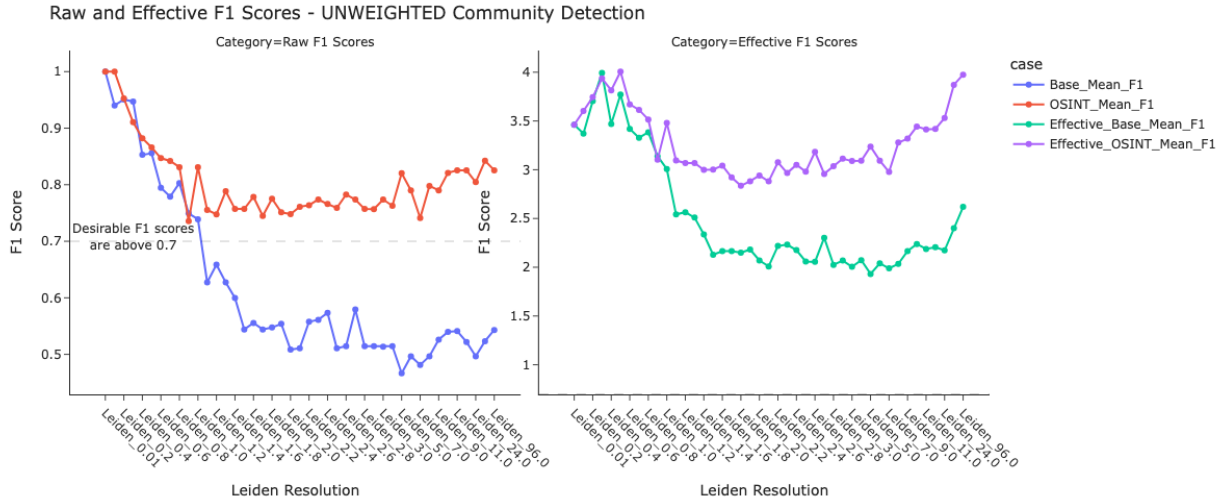
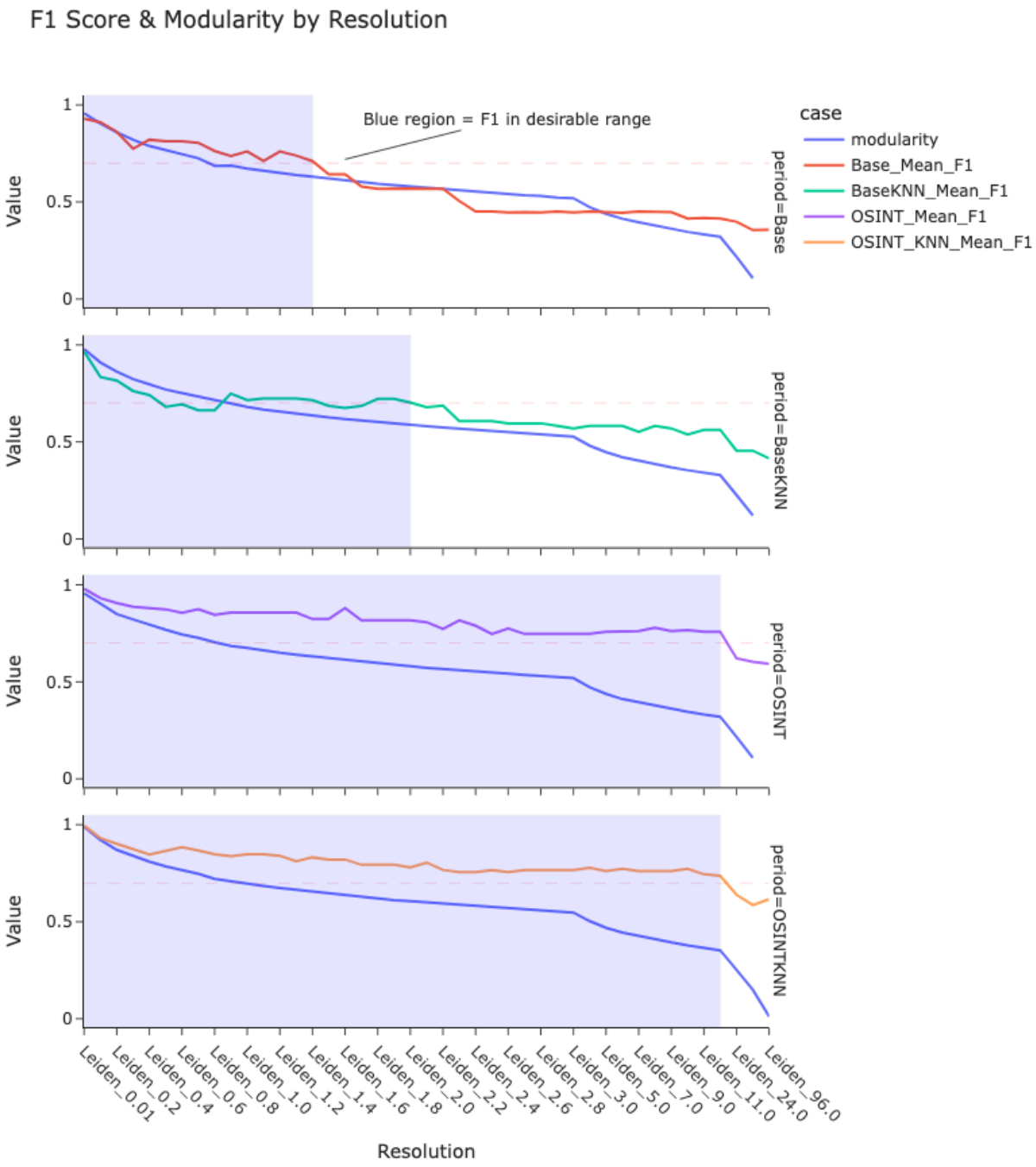


Figure 4-3
Weighted F1 Score and Modularity by Resolution



RQ2. Can we preserve the same level of accuracy by applying reductionism to the OSINT-augmented graphs using fastRP embeddings?

The results for Research Question 2 reveal mixed outcomes regarding the impact of reductionism through FastRP embeddings on fraud ring detection accuracy. While statistically significant improvements were observed across weighted graphs, unweighted graphs exhibited a noticeable decline in performance despite achieving statistical significance.

Weighted Graphs: Reductionism via FastRP embeddings (64-dimensions) significantly improved F1 scores in OSINT-augmented graphs compared to BaseKNN configurations *(W=946, p<.001)*, with a mean difference of *0.147 (SE=0.0073).* The rank-biserial correlation *(r=1.000)* indicates strong directional consistency favoring OSINTKNN over BaseKNN.

Effective F1 scores showed even larger improvements, with a mean difference of *0.5315 (SE=0.0245, W=946, p<.001)*, confirming that reductionism preserves accuracy while adjusting for community count.

Unweighted Graphs: Although statistically significant differences were observed in unweighted graphs (W=903, p<.001), the mean difference of 0.264 (SE=0.0124) was accompanied by a noticeable decline in overall F1 scores compared to weighted graphs (Figure []). Effective F1 scores also showed improvement (mean difference=0.7787, SE=0.0644, W=940, p<.001), but the decline in raw F1 scores suggests that unweighted configurations may be less robust when applying reductionism.

BaseKNN vs OSINTKNN Comparisons: Across all metrics and graph states, OSINTKNN consistently outperformed BaseKNN configurations.

These findings suggest that while reductionism through FastRP embeddings is effective for weighted graphs, unweighted configurations may require additional adjustments to maintain high performance levels.

Table 4-2
Research Question 2 Statistics

BaseKNN vs OSINTKNN Weighted, Unweighted, and Effective F1 Wilcoxon Signed Rank Test

Effective F1 = F1score * log(communityCount)

uw = Unweighted

Paired Samples T-Test

			Statistic	p	Mean difference	SE difference		Effect Size
uw_OSINT_KNN_Mean_F1	uw_BaseKNN_Mean_F1	Wilcoxon W	903 [a]	<.001	0.2640	0.01237	Rank biserial correlation	1.000
uw_OSINT_Effective_Mean_F1	uw_Base_Effective_Mean_F1	Wilcoxon W	940	<.001	0.7787	0.06435	Rank biserial correlation	0.987
OSINT_KNN_Mean_F1	BaseKNN_Mean_F1	Wilcoxon W	946	<.001	0.1472	0.00726	Rank biserial correlation	1.000
OSINTKNN_Effective_Mean_F1	BaseKNN_Effective_Mean_F1	Wilcoxon W	946	<.001	0.5315	0.02487	Rank biserial correlation	1.000
BaseKNN_Mean_F1	uw_BaseKNN_Mean_F1	Wilcoxon W	898 [a]	<.001	0.1906	0.01391	Rank biserial correlation	0.989
OSINT_KNN_Mean_F1	uw_OSINT_KNN_Mean_F1	Wilcoxon W	926	<.001	0.0830	0.00666	Rank biserial correlation	0.958

Note. H_a $\mu_{Measure\ 1 - Measure\ 2} \neq 0$

[a] 1 pair(s) of values were tied

Figure 4-4
F1 Score KNN-derived Edges

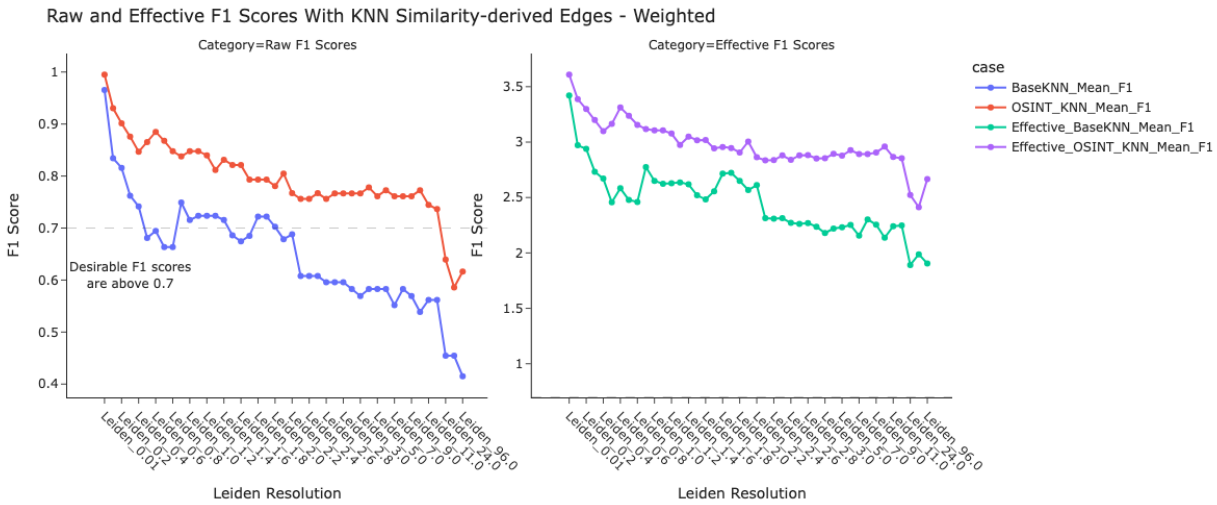
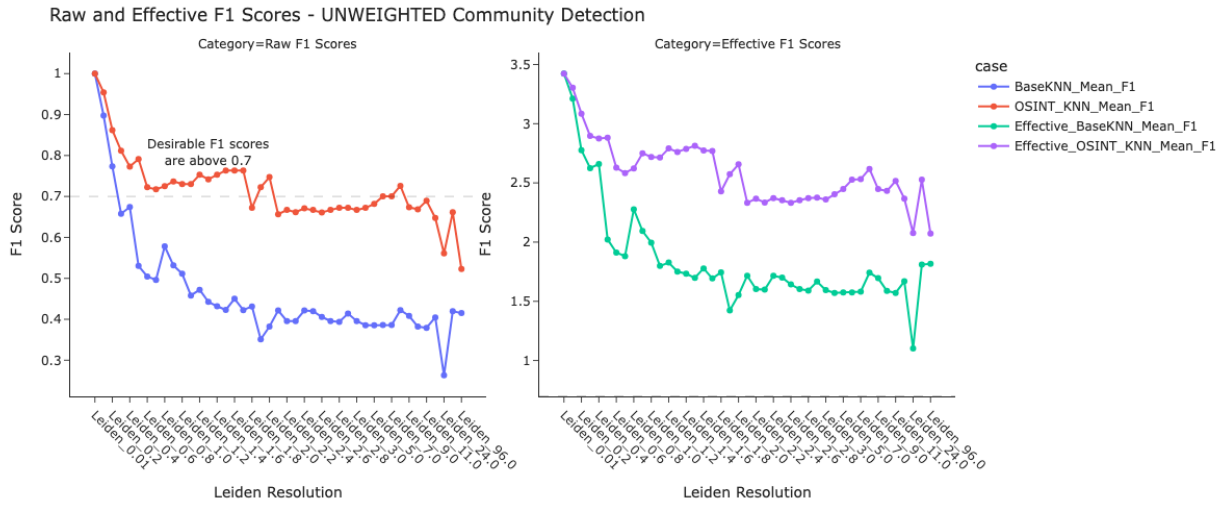


Figure 4-5
F1 Score KNN-derived Edges Unweighted



Case #1 First Choice Laboratory & Sonoran Desert Pathology

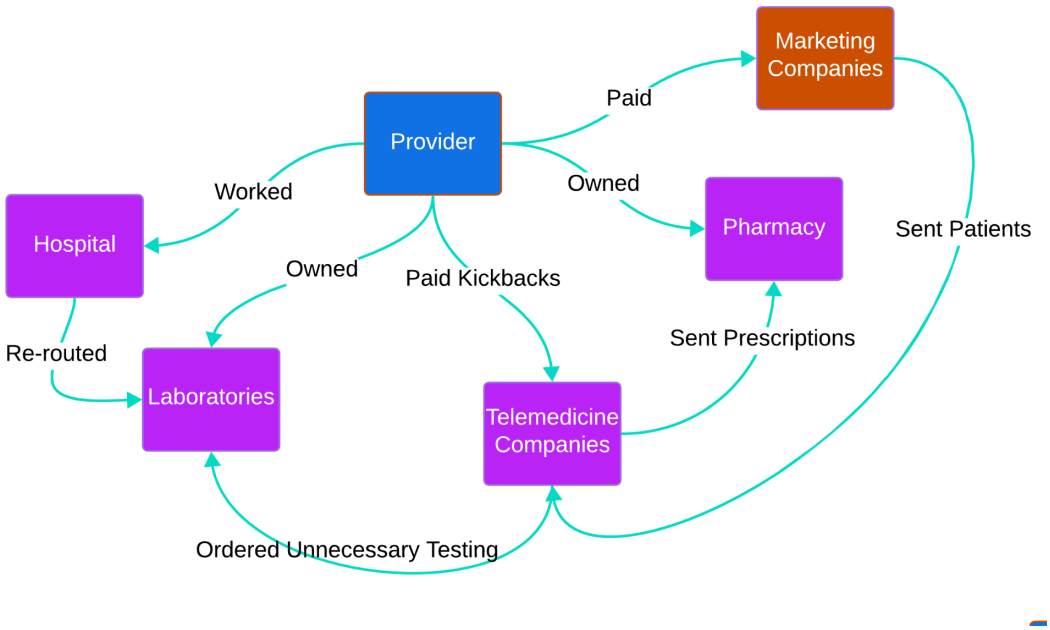
Fraud Type: False Claims + Upcoding + Unnecessary Services + Kickbacks + DME
Total Loss: $97 million
OSINT Sources: Court documents, OpenCorporates, California Business Registry
Pre/Post Node Count: 3 9

This case centers around Daniel Hurt, who orchestrated three Medicare fraud schemes totaling $97 million in losses. Hurt's operations involved kickbacks paid to patient recruiters and telemedicine services to generate medically unnecessary prescriptions. These prescriptions were filled by pharmacies, some owned by Hurt, that billed Medicare for compounded medications at inflated rates. Kickbacks were then funneled through sham contracts with marketers to obscure their illegitimacy. Additionally, Hurt engaged in fraudulent cancer genomic (CGx) testing schemes and misappropriated funds from ECMC-related accounts to pay illegal kickbacks. These activities resulted in significant financial losses to Medicare and exposed vulnerabilities in claims data that lacked connections to intermediaries such as marketers and shell companies (U.S. Department of Justice, 2024).

Scheme Diagram Case #1

The fraud scheme is depicted in Figure [], showing the interactions between actors and their roles in the network. The diagram highlights Hurt's coordination between pharmacies, marketers, patient recruiters, and telemedicine services.

Figure 4-6
Case #1 Case Scheme Diagram



Egonet for Case #1

The initial graph contained three nodes corresponding to providers directly involved in the scheme: First Choice Laboratory, Sonoran Desert Pathology Associates, and Landmark Diagnostics. Post-OSINT augmentation revealed six additional nodes linked to fraudulent activity through shared addresses and business associations. These new nodes included Stephen Sinkoe (early co-conspirator), Robere Missirian (officer of Sonoran Desert Pathology), and Analyze Pathology Inc., among others. Figure 4-8 shows the shared connections between Sonoran Desert Pathology and First Choice. These shared connections are reflected in the Jaccard similarity metrics.

Table 4-10
Case #1 Nodes

Node	Type	NPI#
First Choice Laboratory	Provider	1790109742

Sonoran Desert Pathology Associates	Provider	1205162955
Landmark Diagnostics	Provider	1699172866
Stephen Sinkoe	Provider	1992811996
Stephen m. Sinkoe dpm	Supplier	1376712208
Robere Missirian	Provider	1326255134

Figure 4-7
Case #1 Supplier/Provider Nodes

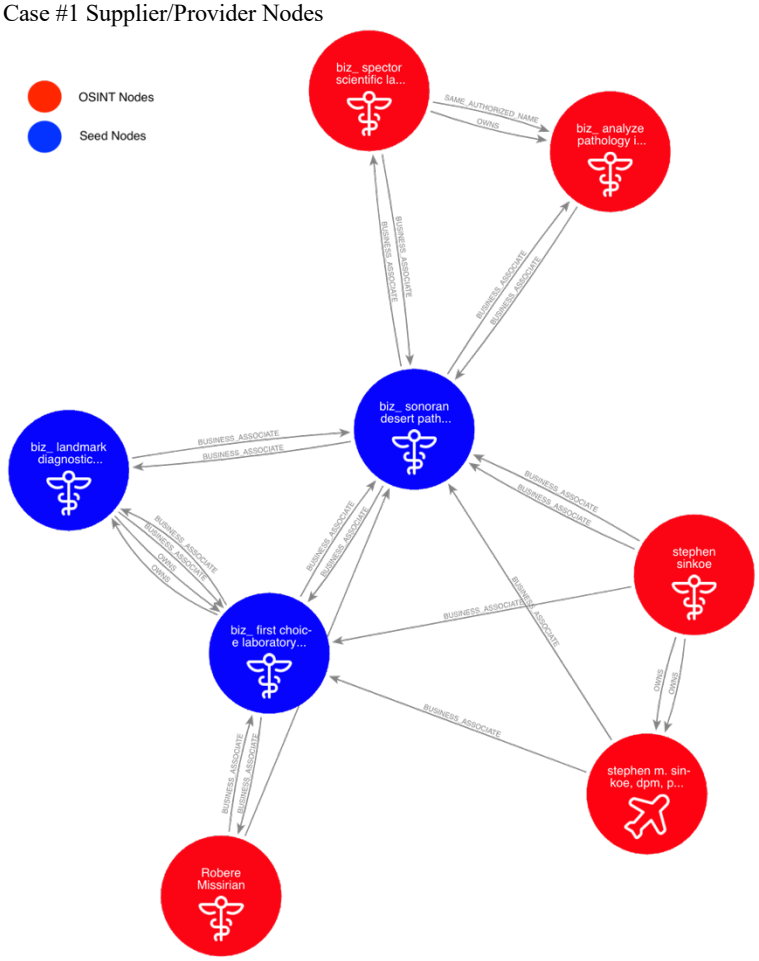
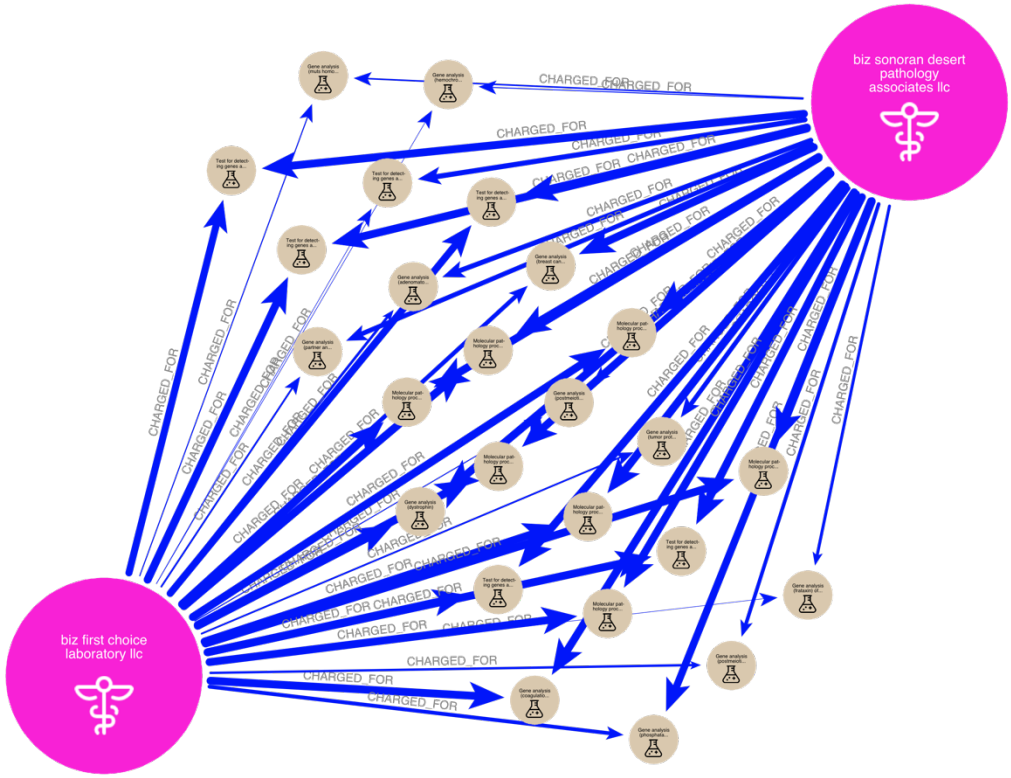


Figure 4-8
Shared Connections Case #1



Case # 1 OSINT

Open Source Intelligence (OSINT) investigations were conducted with the aim of locating associated entities possessing National Provider Identifier (NPI) numbers that might facilitate billing to Medicare and possibly contribute to the continuation of illicit activities. Referenced court documents are included in the Appendix.

Table 11 Case #1 OSINT

Table 4-11
Case #1 OSINT

OSINT Name	OSINT NPI	OSINT SOURCE	SUMMARY
Stephen Sinkoe	1992811996	Court documents	Stephen was an early co-conspirator who was first in the door to cooperate with Federal authorities. His plea agreement

OSINT Name	OSINT NPI	OSINT SOURCE	SUMMARY
			created a domino effect that led to the quick prosecution of the other co-defendants in his case.
Stephen m. Sinkoe dpm	1376712208	OpenCorporates	Business entity under Sinkoe
Robere Missirian	1326255134	California business registration	Signer/officer of Sonoran Desert Pathology
Analyze Pathology Inc	1396384251	Open Corporates	Whistleblower in case #1 has businesses registered to the same address as the owner of Analyze Pathology Inc. Both whistleblower and owner are co-owners of Titus AP Partnership
Spector Scientific Inc	1215390422	OpenCorporates	Same owner and address as Analyze Pathology Inc

Jaccard Similarity

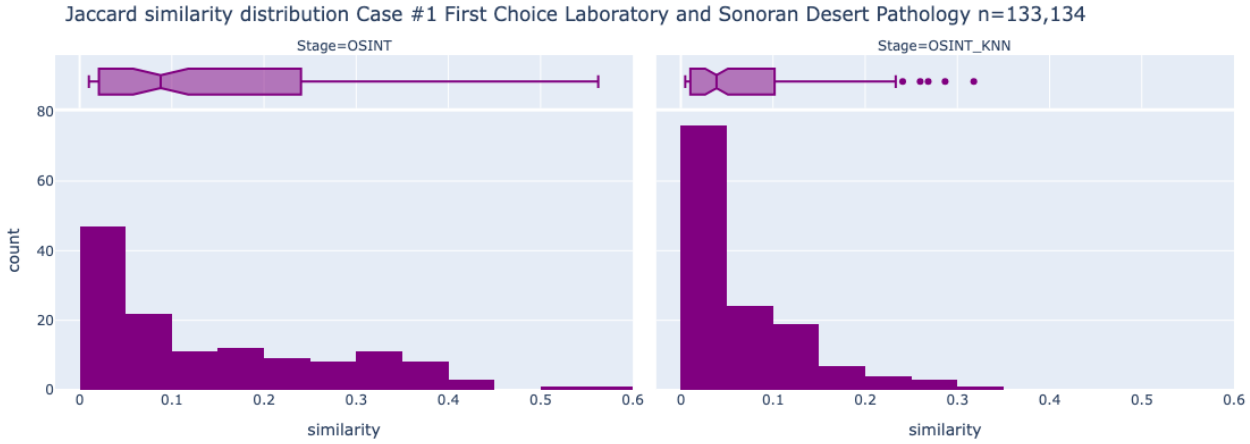
Post-augmentation Jaccard mean and median similarity scores decreased substantially.

Table 4-12
Case #1 Jaccard Similarity Table

Stage	Count	Mean	Std	Min	25%	50%	75%	Max
Base	7.00	0.17	0.14	0.08	0.09	0.10	0.24	0.38
BaseKNN	7.00	0.11	0.18	0.01	0.03	0.04	0.08	0.52
OSINT	21.00	0.11	0.12	0.01	0.04	0.06	0.10	0.41
OSINT_KNN	21.00	0.05	0.06	0.01	0.01	0.02	0.05	0.27

Figure 4-9
Case #1 Jaccard Similarity

Figure 4-10
Case #1 Jaccard Similarity Distribution Pre-OSINT



Neighborhood Statistics Case #1

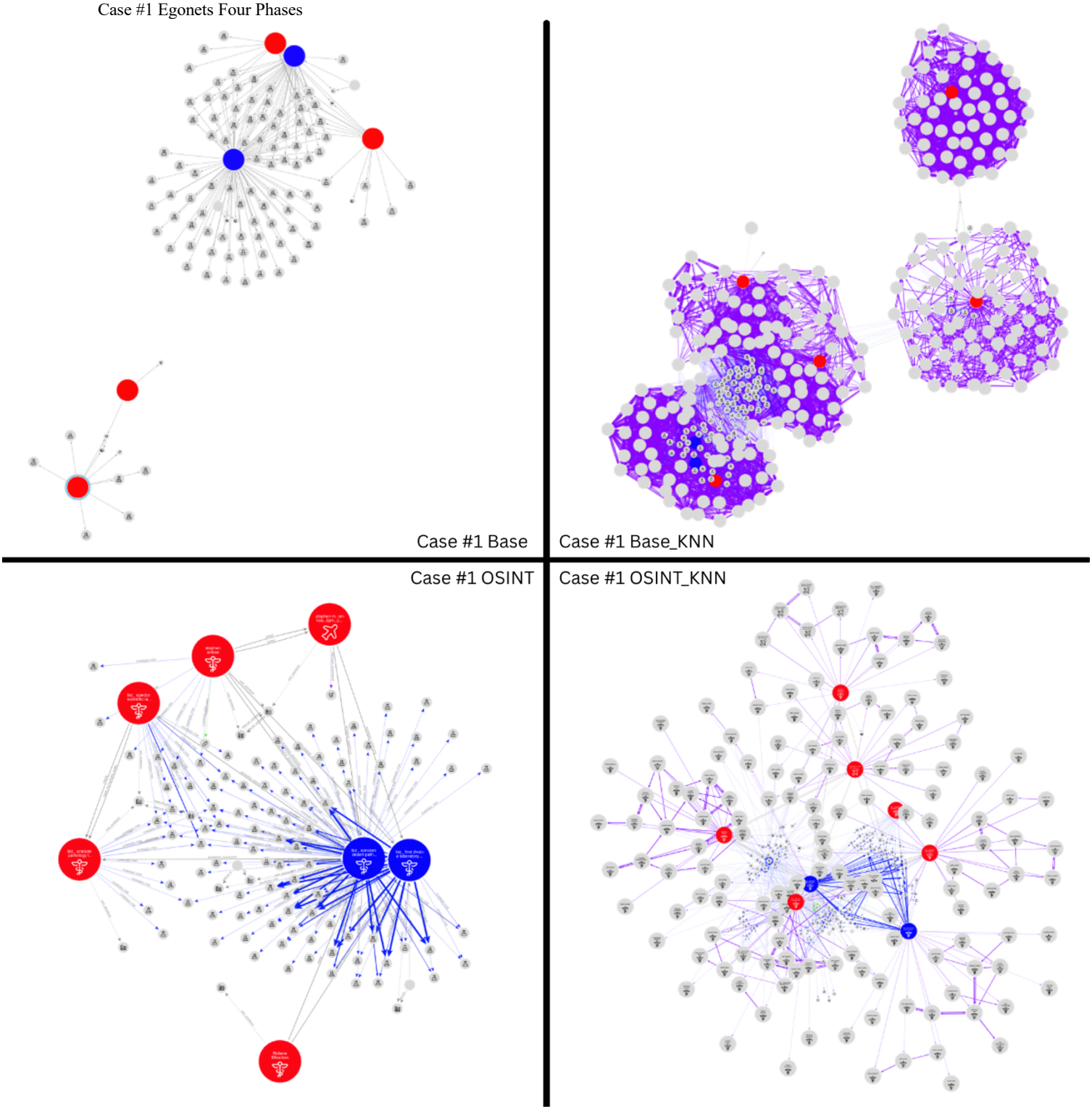
Table 4-13
Case #1 Jaccard Similarity Distribution Pre-OSINT

Neighborhood Metrics Case #1				
	Base	BaseKNN	OSINT	OSINT_KNN
Density	0.02	0.09	0.03	0.02
Average Degree	4.34	30.95	2.97	5.16
Average Weighted Degree	339,689	2,976,671	2,630,533	1,308,667
Clustering Coefficient	0.13	0.55	0.39	0.37
Diameter	8.00	8.00	4.00	4.00
Average Path Length	3.28	3.47	2.30	2.76

Key Observations:

- **Density:** Increased marginally post-OSINT, but remained unchanged with the KNN-derived edges.
- **Clustering Coefficient:** Increased significantly post-OSINT (from 0.13 to 0.39).
- **Diameter:** Reduced from 8 to 4 post-OSINT augmentation, reflecting a more compact network structure.
- **Weighted Degree:** Revealed substantially more economic activity post-OSINT.

Figure 4-11
Case #1 Egonets Four Phases



Case #2 J. Stanton

Fraud Type: False Claims + Upcoding + Unnecessary Services + Kickbacks

Total Loss: Dollar amount not specified. 22kg of narcotic prescriptions across 21 patients. 2 doctors sentenced to 17 years in prison. $1.5 million judgement against 1 doctor.

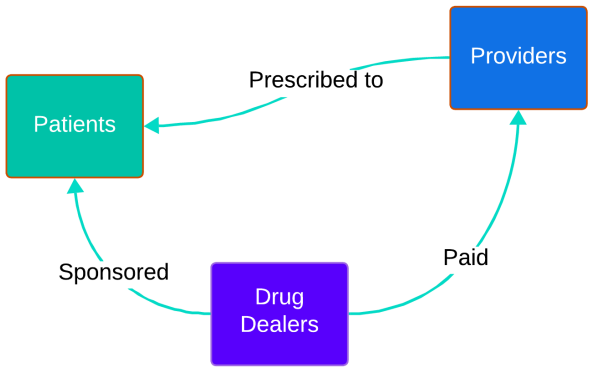
OSINT Sources: Court documents, OpenCorporates, California Business Registry, TN Medical Board

Pre/Post Node Count: 2, 20

On April 13, 2023, Dr. John Stanton was sentenced to 120 months in prison for his role in conspiring to unlawfully distribute controlled substances. Stanton was the medical director for Gateway Medical Associates, a pain management clinic owned by Dr. John Maccarone. In 2016, Tennessee began requiring pain clinics to employ medical directors. Dr. Maccarone lacked the required certificate in pain management, and thus paid Dr. Stanton, who practiced orthopedic surgery next door, $1,500 per week to sign as the acting medical director for Gateway Medical Associates. According to the indictment, patients would form long lines at the clinic and some traveled more than 10 hours for their appointment. The high-volume of narcotic prescriptions despite patients failing drug tests earned Gateway Medical Associates the reputation of being a pill mill. Two of Stanton and Maccarone's codefendants would 'sponsor' patients to receive prescriptions so that the drugs could be diverted for street sale (U.S. Department of Justice, 2023).

Scheme Diagram Case #2

Figure 4-12
Case #2 Fraud Scheme Diagram



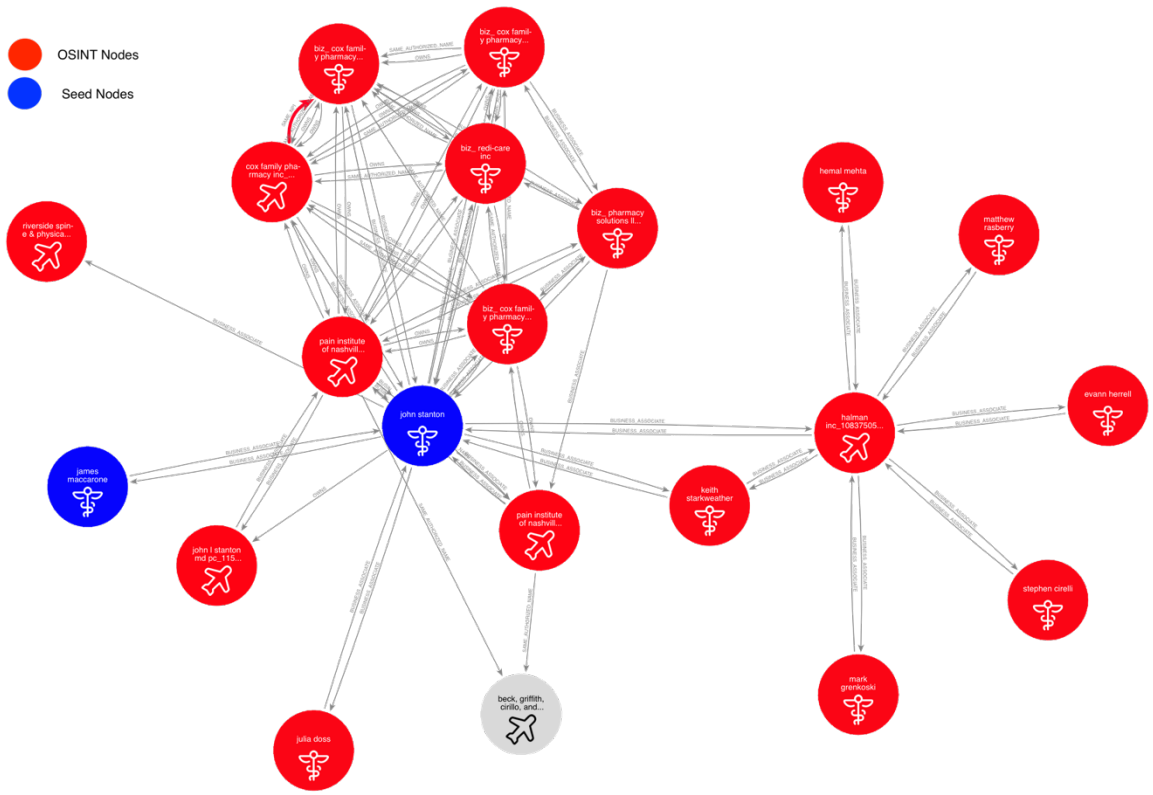
Egonet for Case #2

The initial graph contained two nodes of providers directly involved in the scheme. Supervisory relationships noted on the Tennessee Medical Board records site and court documents revealed several more co-conspirators and associates involved in similar activity.

Table 4-14
Case #2 Nodes

Node	Type	NPI#
James Maccarone	Provider	1861590002
John Stanton	Provider	1245233063
Clarksville Pain Institute	Supplier	1669749537
Joint and Spine Pain Center	Supplier	1154386282
White House Pain Institute	Supplier	1356773626
Cox Family Pharmacy	Supplier	1538170030
Keith D StarkWeather	Provider	1407859598
Debbie Cox	Provider	1215016860
Cox Family Pharmacy	Supplier	1316135536
Cox Family Pharmacy	Supplier	1043617269
Halman	Supplier	1083750574
Hemal Mehta	Provider	1174533772
Matthew Rasberry	Provider	1154581825
Mark Grenkoski	Provider	1174567754
Stephen Cirelli	Provider	1134120777
Evann Herrell	Provider	1306850466

Figure 4-13
Case #2 Supplier/Provider Nodes



Case #2 OSINT

Table 4-15
Case #2 OSINT

OSINT Name	OSINT NPI	OSINT SOURCE	SUMMARY
Debbie Cox	1215016860	TN Medical Board	Supervised by Dr. John Stanton
Cox Family Pharmacy	1316135536 1043617269	Prior Indictment	Indictment against Debbie and Michael Cox – unknown prior to case selection
Pain Institute of Nashville	1033694724 1073049102	Prior Indictment	Owned by Michael Cox
Riverside spine & physical medicine, pc	1720407570	Vitals.com and court documents	Dr. Stanton was affiliated with this clinic
Medsource scripts	1710719380	Prior Indictment	One of several pharmacies operating out of the same location as Cox Pharmacy
Joint and Spine Pain Center	1154386282	Webmd.com and court documents	Dr. Stanton affiliated with this clinic

OSINT Name	OSINT NPI	OSINT SOURCE	SUMMARY
Halman	1083750574	Court documents and OpenCorporates	A confidential informant provided evidence of collusion between the owner of Halman and Dr. Stanton
Keith D Starkweather	1407859598	OpenCorporates	Registered agent/owner of Bone and Joint Group
Bone and Joint Group	NA	Webmd.com and court documents	Dr. Stanton affiliated with this clinic
Matthew Raspberry	1154581825	Indictment against Halman	Co-conspirator in Halman case
Mark Grenkoski	1174567754	Indictment against Halman	Co-conspirator in Halman case
Stephen Cirelli	1134120777	Indictment against Halman	Co-conspirator in Halman case
Evann Herrell	1306850466	Indictment against Halman	Co-conspirator in Halman case
Michael Cox	NA	Indictment and OpenCorporates	Married to Debbie Cox. Authorized name/owner of Cox Family Pharmacy
Redi-care Inc	1396835716	OpenCorporates	Found incidentally in graph and confirmed through OpenCorporates Company is owned by Debbie and Michael Cox
Pharmacy Solutions	1295274306	Court documents	Operating out of same location as Cox family pharmacy. Owned by J. Prichard aka Medsource
Julia Doss	1093340804	OpenCorporates	Same address as Joint and Spine Pain Center. Julia Doss listed as owner

Jaccard Similarity

Figure 4-14
Jaccard Similarity Case #2

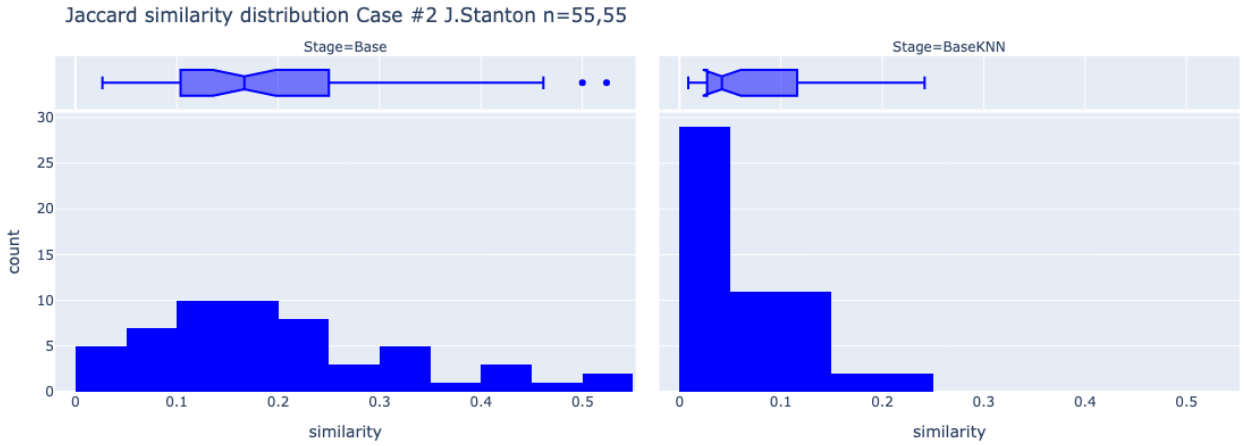
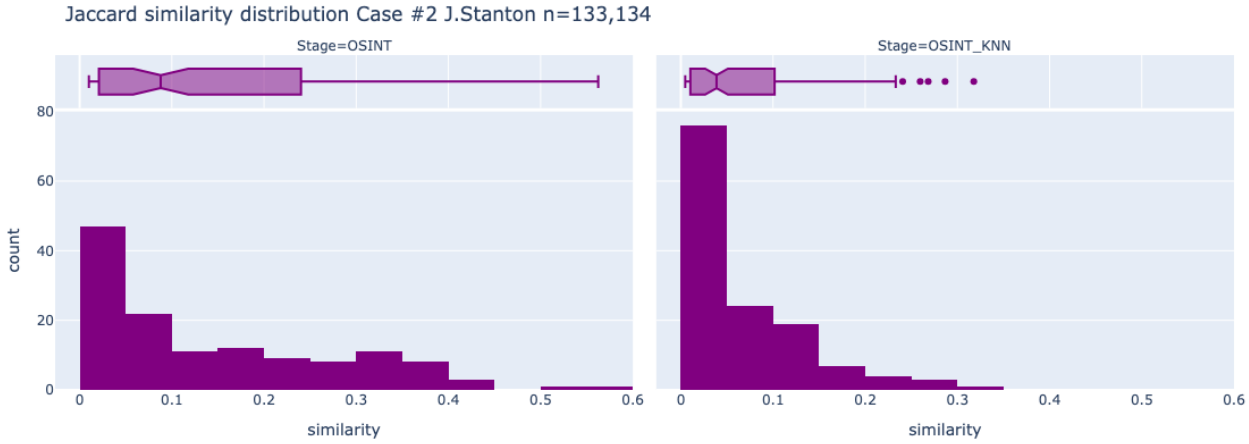


Figure 4-15
Case #2 Jaccard Similarity post-OSINT



Neighborhood Statistics Case #2

Key Observations

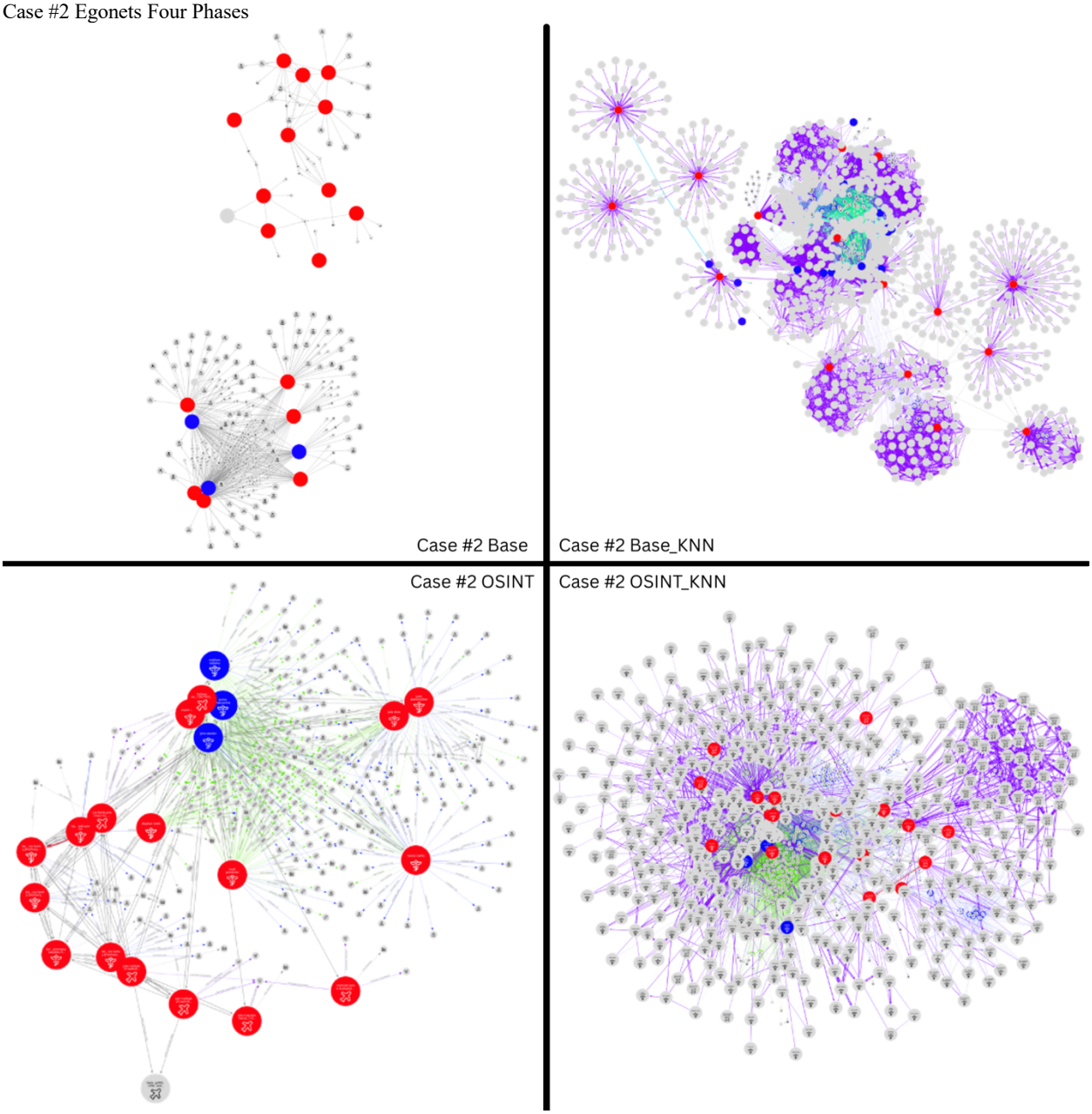
- **Clustering Coefficient:** Increased dramatically post-OSINT
- **Diameter:** Reduced from 7 to 5 with OSINT, indicating more direct connections between conspirators.

- **Weighted Degree:** OSINT augmentation revealed financial interactions over 20 times larger than those visible in claims data alone, highlighting the extensive economic scale of the fraud.

Table 4-16
Case #2 Neighborhood Statistics

Neighborhood Statistics Case #2				
	Base	BaseKNN	OSINT	OSINT_KNN
Density	0.02	0.03	0.02	0.02
Average Degree	4.02	35.04	4.22	18.39
Average Weighted Degree	33,150.34	274,273.38	49,596.36	98,184.04
Clustering Coefficient	0.02	0.45	0.10	0.21
Diameter	12.00	7.00	5.00	5.00
Average Path Length	3.21	3.37	3.20	2.89

Figure 4-16
Case #2 Egonets Four Phases



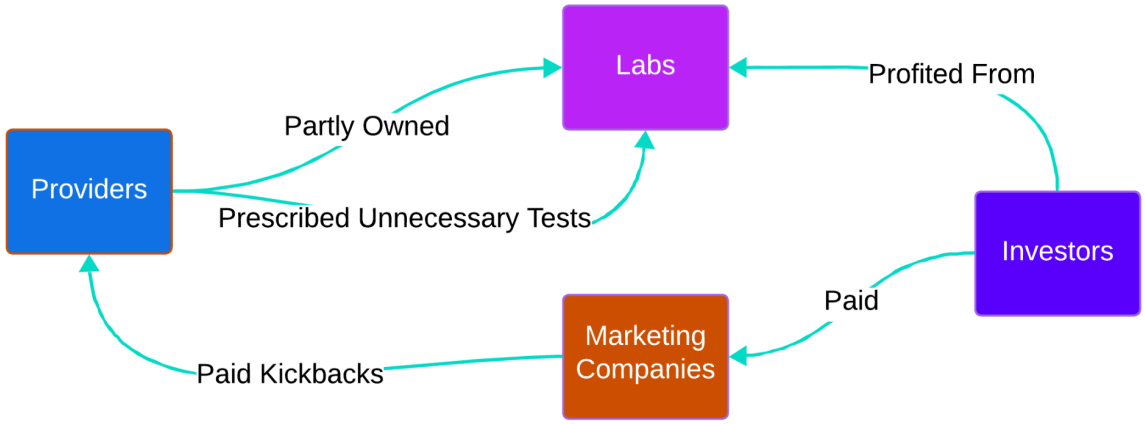
Case #3 Canova

Fraud Type: Upcoding + Unnecessary Services + Kickbacks
Total Loss: $300 million
OSINT Sources: Court documents, OpenCorporates
Pre/Post Node Count: 2 11

In April 2022, 11 defendants pled guilty to charges stemming from a $300 million Medicare fraud scheme. Two doctors and a nurse practitioner received illegal kickbacks in exchange for ordering medically unnecessary lab tests. The lab owners used marketing companies to disburse the kickbacks as consulting fees. Knowing that kickbacks could be more easily disguised in a provider-ownership model, physicians were offered ownership stakes in Reliable Laboratories LLC contingent on providers meeting a set quota for lab test referrals (United States vs. Madison).

Scheme Diagram Case #3

Figure 4-17
Case #3 Scheme Diagram



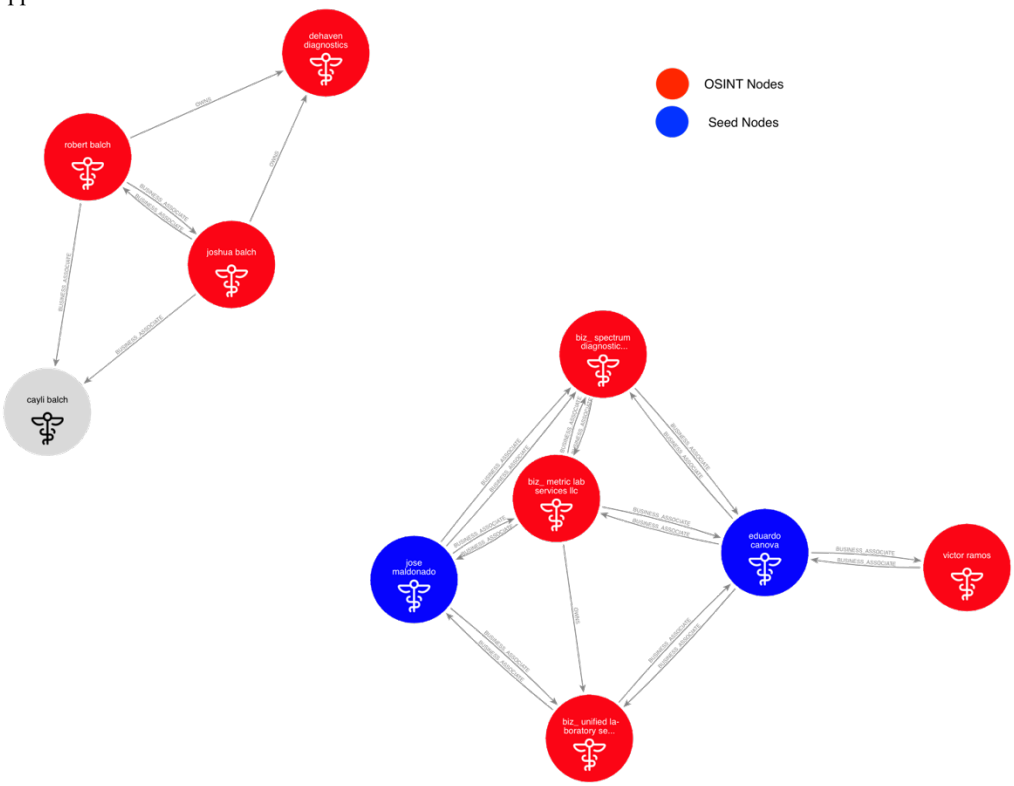
Egonet for Case #3

Table 4-17
Case #3 Egonet

Node	Type	NPI#
jose maldonado	Provider	1417091919
eduardo canova	Provider	1376854968
reliable labs	Supplier	1316326838
spectrum diagnostic	Supplier	1497173348

unified laboratory services	Supplier	1841546678
keith allen wichinski	Provider	1306046644
DeHaven Diagnostics	Supplier	1124369459
Star Labs	Supplier	1952843294
Metric Labs LLC	Provider	1699287300
Victor Ramos	Provider	1720587595
Robert Balch	Provider	1477751774

Figure 4-18
Case #3 Supplier/Provider Nodes



Case #3 OSINT

Table 4-18
Case #3 OSINT

OSINT Name	OSINT NPI	OSINT SOURCE	SUMMARY
Reliable Labs		Court documents	Co-defendant in kickbacks case
Spectrum Diagnostic	1497173348	Court documents	Co-defendant in kickbacks case

Unified Laboratory Services	1841546678	Court documents	Co-defendant in kickbacks case
Keth Allen Wichinski		Court documents	Co-defendant in kickbacks case
DeHaven Diagnostics	1124369459	OpenCorporates	Physician-owned lab registered to the same address as unified laboratory services
Star Labs	1952843294	OpenCorporates	Registered to the same owner as Unified Laboratory Services. Dissolved right before Unified formation.
Metric Labs LLC	1699287300	OpenCoporates	Co-founder of Unified and Reliable Labs (Sherman Kennereson) owns Metric Labs
Victor Ramos	1720587595	OpenCorporates	Doctor with address at Canova's practice address
Robert Balch	1477751774	OpenCorporates	Owner of DeHavenDiagnostics

Jaccard Similarity Case #3

Figure 4-19
Case #3 Jaccard Similarity Distribution

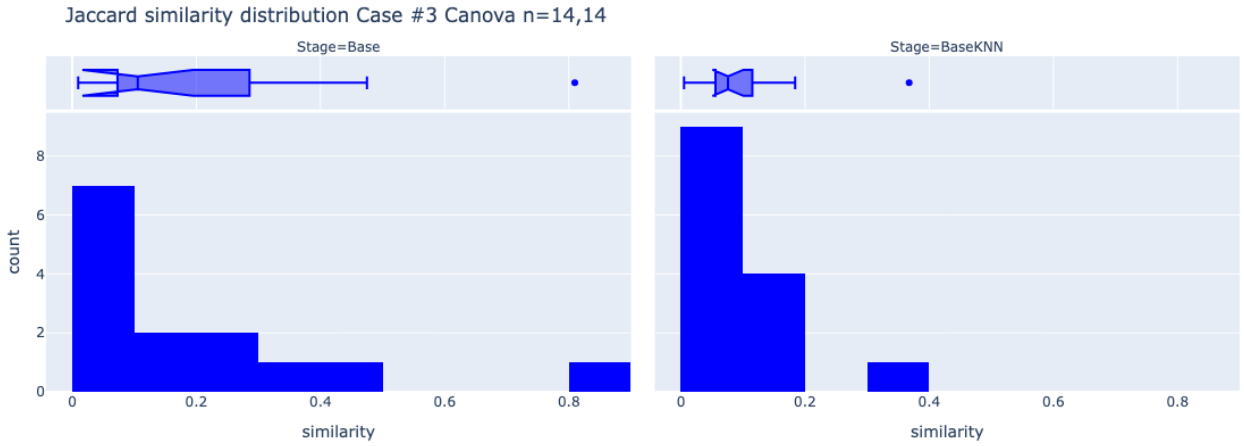
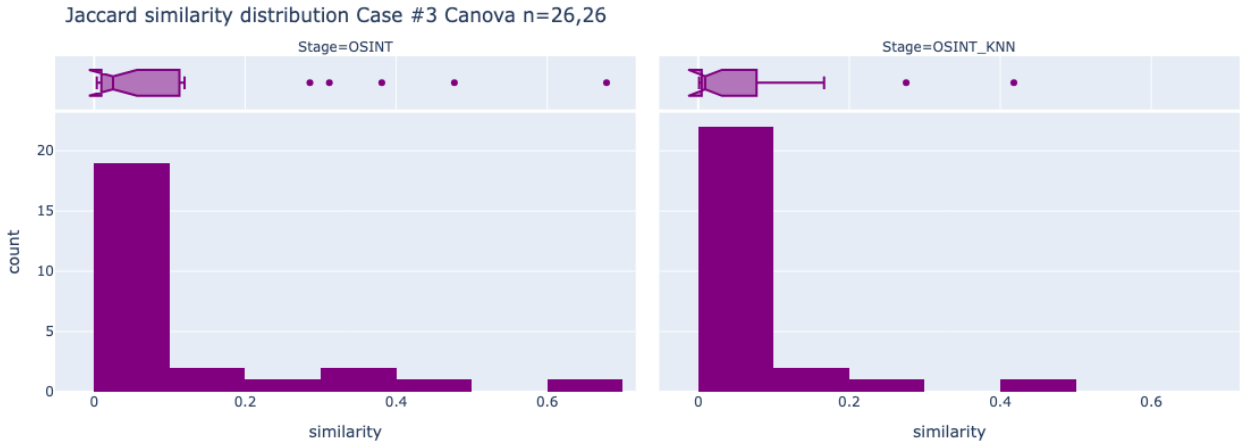


Figure 4-20
Case #3 Jaccard Similarity Pre-OSINT

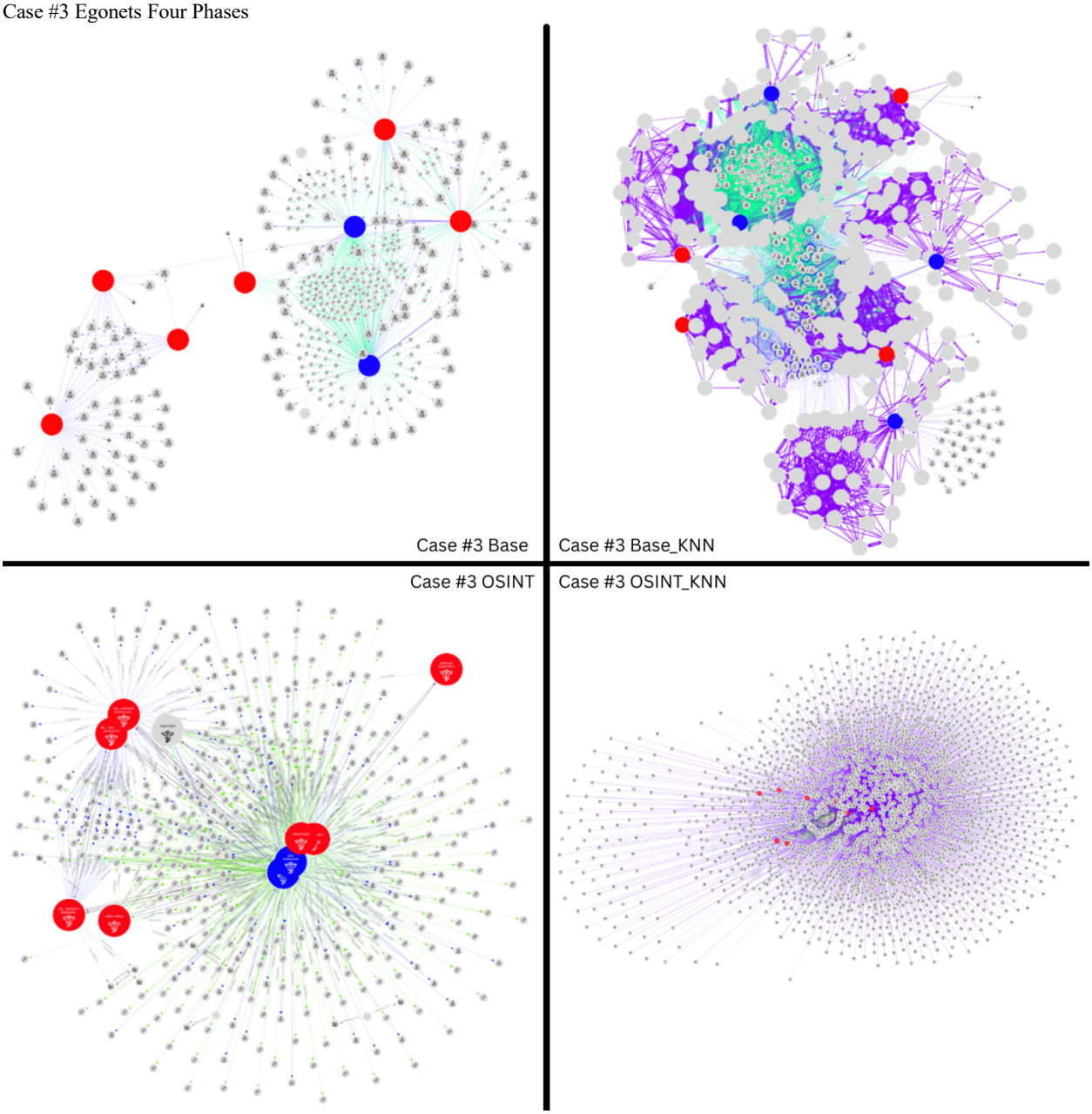


Neighborhood Statistics Case #3

Table 4-19
Case #3 Neighborhood Statistics

Neighborhood Metrics Case #3				
	Base	BaseKNN	OSINT	OSINT_KNN
Density	0.01	0.06	0.01	0.01
Average Degree	3.31	42.56	3.57	16.01
Average Weighted Degree	45,063.69	360,095.87	52,990.82	45,034.74
Clustering Coefficient	0.00	0.20	0.13	0.35
Diameter	8.00	5.00	5.00	5.00
Average Path Length	3.71	2.64	2.77	2.29

Figure 4-21
Case #3 Egonets Four Phases



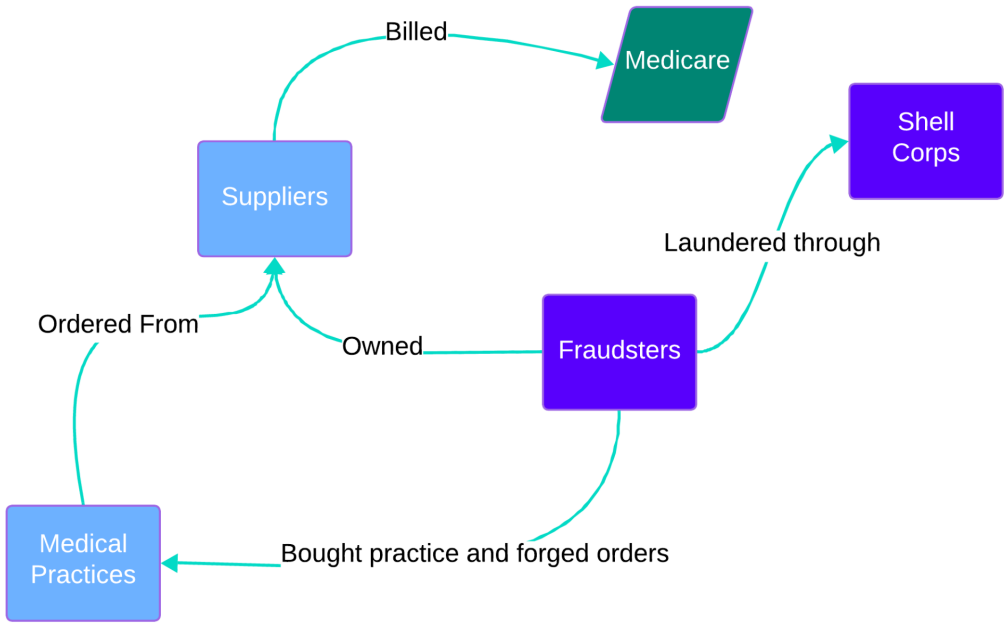
Case #4 DME Bust Out

Fraud Type: False Claims + Unnecessary Services + DME + Identity Theft
Total Loss: $18 million to Medicare. $231 million in gold, luxury vehicles, and other assets seized.
OSINT Sources: Court documents, OpenCorporates
Pre/Post Node Count: 10 34

In June 2024, a dozen individuals were indicted for their roles in an $18 million DME bust out scheme. As part of the scheme, the defendants purchased existing health care practices to obtain patients' insurance information and providers' credentials. The defendants falsified orders for DME through companies they controlled. The Medicare reimbursements were then quickly transferred to several shell companies and subsequently to 35 co-conspirators who cashed checks from the shell corps in amounts small enough to avoid reporting under American money-laundering (AML) laws (United States vs. Izquierdo).

Scheme Diagram Case #4

Figure 4-22
Case #4 Scheme Diagram



Egonet Case #4

Table 4-20
Case #4 Egonet Nodes

Node	Type	NPI#
one way medical llc_1184234866	Supplier	1184234866
del prado medical supply inc_1235716606	Supplier	1235716606
community medical supply inc_1235737321	Supplier	1235737321
one sky medical llc_1285237842	Supplier	1285237842
cleveland medical supply inc_1336758945	Supplier	1336758945
new way med supply llc_1467047233	Supplier	1467047233
premiere medical supply inc_1548879141	Supplier	1548879141
pegasus medical supply llc_1588262828	Supplier	1588262828
eddie's med supply llc_1891399408	Supplier	1891399408
medica merica supplies corp_1972103380	Supplier	1972103380
alpha medical consulting inc	Supplier	1043878424
lazarus services llc	Supplier	1366892242
p medical supplies	Supplier	1730123092
minha medical supplies	Supplier	1215712120
md healthcare	Supplier	1679172746
better choice medical supply inc	Supplier	1861086399
elegance medical supply	Supplier	1487306155
ysys m lastres	Provider	1912597717
neisy abreu jimenez	Provider	1821501487
harmony home care solutions llc	Supplier	1033710249
biz boca toxicology, llc	Provider	1811360977
med medical supplies corp_1013590371	Supplier	1013590371
one way medical llc_1184234866	Supplier	1184234866
del prado medical supply inc_1235716606	Supplier	1235716606
community medical supply inc_1235737321	Supplier	1235737321
1 med supplier corp_1245851302	Supplier	1245851302
one sky medical llc_1285237842	Supplier	1285237842
cleveland medical supply inc_1336758945	Supplier	1336758945
new way med supply llc_1467047233	Supplier	1467047233
premiere medical supply inc_1548879141	Supplier	1548879141
pegasus medical supply llc_1588262828	Supplier	1588262828
better choice medical supply inc_1861086399	Supplier	1861086399
eddie's med supply llc_1891399408	Supplier	1891399408
medica merica supplies corp_1972103380	Supplier	1972103380

Figure 4-23
Case #4 Supplier/Provider Nodes

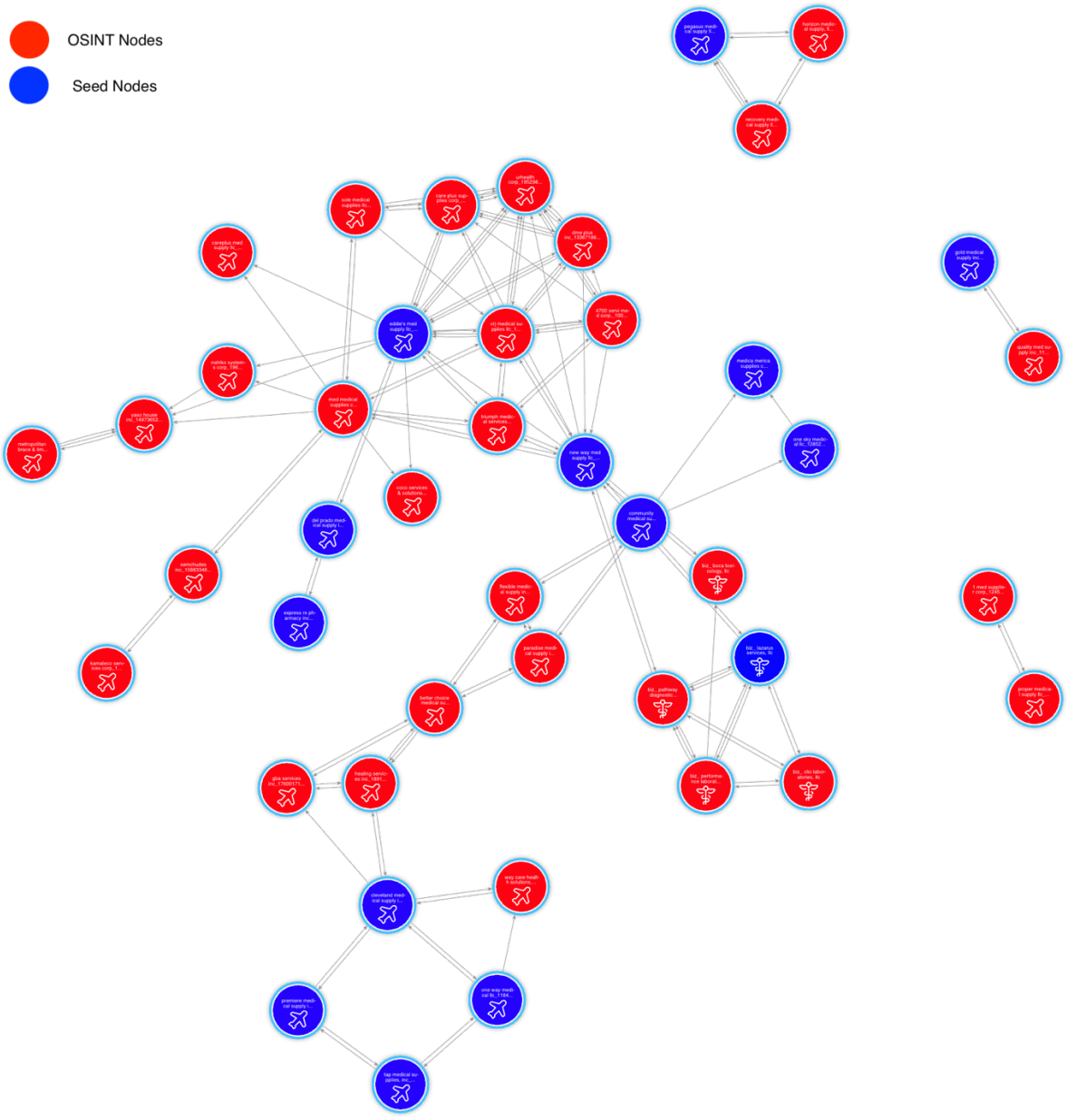


Table 4-21
Case #4 OSINT

OSINT Name	OSINT NPI	OSINT SOURCE	SUMMARY
Performance Laboratories	1245657675	OpenCorporates	Same registered owner as Lazarus Services LLC (Mondelli case linked to address for 1 med)

OSINT Name	OSINT NPI	OSINT SOURCE	SUMMARY
Pathway Diagnostics	1710347703	OpenCorporates	Same registered owner as Lazarus Services LLC (Mondelli case)
Proper Medical Supply	1508474180	OpenCorporates Google Reviews	Same owner as 1 med. Google reviews mentions fraudulent activity
MD Healthcare LLC	1679172746	OpenCorporates	Same owner as codefendant Caleb Espinoza
Harmony Home Care Solutions	1033710249	OpenCorporates	Same owner as Cleveland Medical Supply
Minha Medical Supplies	1215712120	OpenCorporates	Linked to same address as New Way Med Supply (defendant)
Triumph Medical Services Corp	1972124030	OpenCorporates	Owned by Caleb Espinoza
Elegance Medical Supply	1487306155	OpenCorporates	Same registered owner as Del Prado Medical Supply
DMS Wholesale		OpenCorporates	Same owner as Cleveland Medical Supply
NEISY ABREU JIMENEZ, RBT		NPIDB.org	Same registered address as DMS Wholesale
Ysys M Lastres	1912597717	NPIDB.org	Same registered address as Premiere Medical (this is an apartment building)
Boca Toxicology	1811360977	OpenCorporates	Address for 1 med supply was used by another individual Michael Mondelli charged in a separate fraud ring case
Clio Laboratories	1013382431	OpenCorporates	Mondelli case
Alpha Medical Consulting	1043878424	OpenCorporates	Mondelli case
Lazarus Services LLC	1366892242	OpenCorporates	Mondelli case
P Medical Supplies	1730123092	OpenCorporates	Owned by defendant Pedro Perez
Better Choice Medical Supply	1861086399	OpenCorporates	Same address as Cleveland Medical Supply
Gonzalez Rehab Professionals	1912777657	OpenCorporates	Same address as Cleveland Medical Supply
1 Med Supplier Corp	1245851302	OpenCorporates	Same address as New Way Medical (defendant)

OSINT Name	OSINT NPI	OSINT SOURCE	SUMMARY
Quality Med Supply Inc	1154889970	OpenCorporates	Shares address with Gold Medical. Found incidentally by address on graph.
Med Medical Supplies	1013590371	OpenCorporates	Owned by Caleb Espinoza
Crj Medical Supplies	1780207142	OpenCorporates	Authorized name is Caleb Espinoza but listed owner is Ruber A Lambertt
4700 Servi Med Corp	1003470600	Graph	Caleb listed as president. Corp owned by Jesse
Sole Medical Supplies	1932780665	OpenCorporates	Same owner as Urhealth Corp
Urhealth Corp	1952981698	OpenCorporates	Caleb is authorized signer. This owner also owns Sole Medical Supplies
Care Plus Supplies	1417552456		
GBA Services Inc	1760017156	FLAACO	Shares address with Cleveland Medical (defendant)
Semchudes Inc	1588334684	FLAACO	Same owner as Med Medical
Way Care Health Solutions	1043804412	FLAACO	Same address as One Way Medical and last name of Cleveland owner
DME Plus	1336718642	FLAACO OpenCorporates	Caleb is authorized signer
Kamaleco Services Corp	1083386148	FLAACO	Linked to owner address of Semchudes which is owned by Med Medical owner
Nahiko Systems Corp	1962000208	FLAACO	Next door to Eddie's Medical Supply and Med Medical Supply
Yaso House	1497365225	FLAACO	Same complex as Eddie's Medical and Med Medical Supply
Metropolitan Brace & Limb		Graph	Same signer as Yaso House and same address
Coco services & Solutions Inc	1891392072	FLAACO	Same complex as Eddie's and Med Medical
Careplus Med Supply LLC	1053999888	FLAACO	Same complex as Eddie's and Med Medical
Horizon Medical	1679059810	FLAACO Court documents	Same address as authorized signer for Pegasus Medical Supply. This company is part of a third/separate indictment.

OSINT Name	OSINT NPI	OSINT SOURCE	SUMMARY
Healing Services	1891327839	FLAACO	Same signer as Better Choice Medical. Shared address with Cleveland Medical
Flexible Medical Supply	1841879897	FLAACO	Same complex as Better Choice Medical Supply and Community Choice Medical
Paradise Medical Supply	1477019776	FLAACO	Same complex as Community Medical and Better Choice Medical

Jaccard Similarity Case #4

Figure 4-24
Case #4 Jaccard Similarity

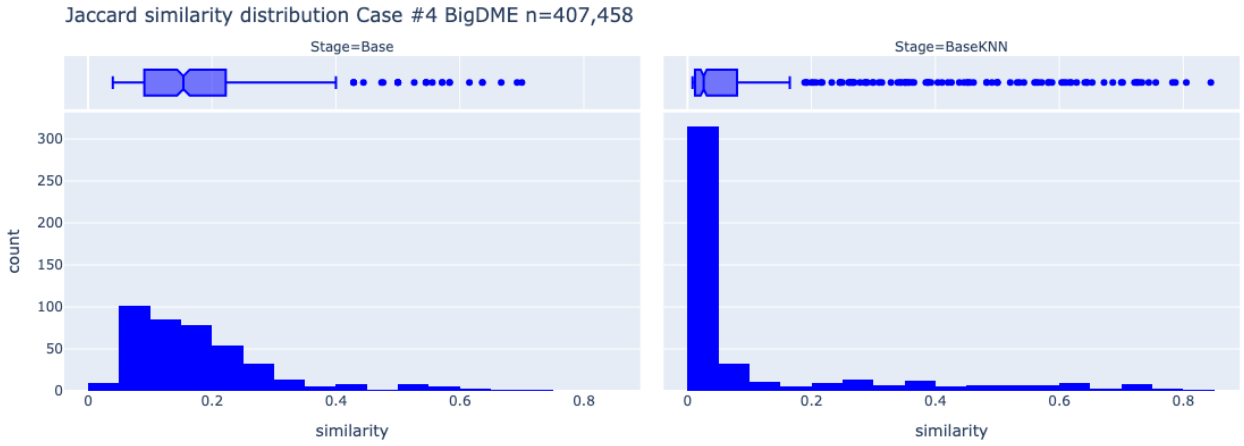
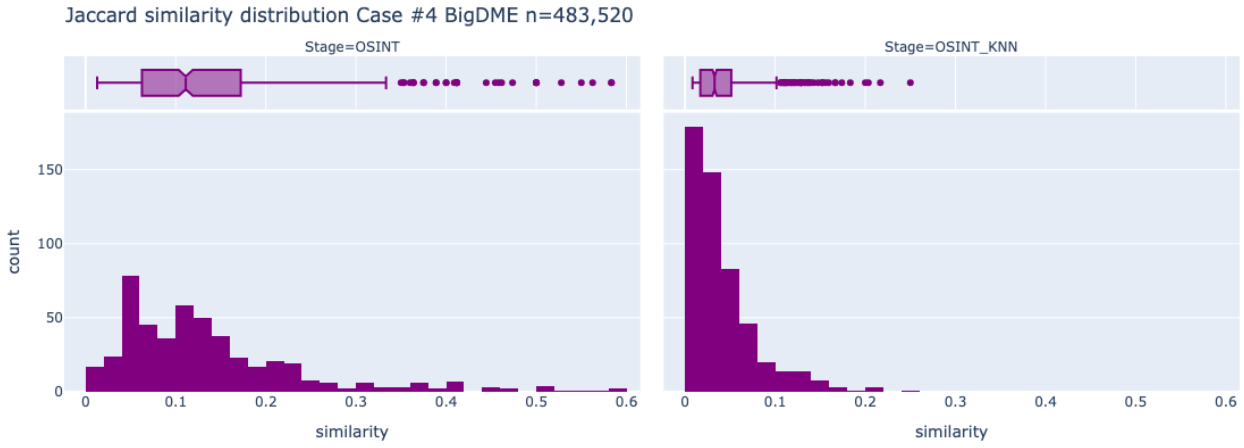


Figure 4-25
Case #4 Jaccard Similarity Pre-OSINT

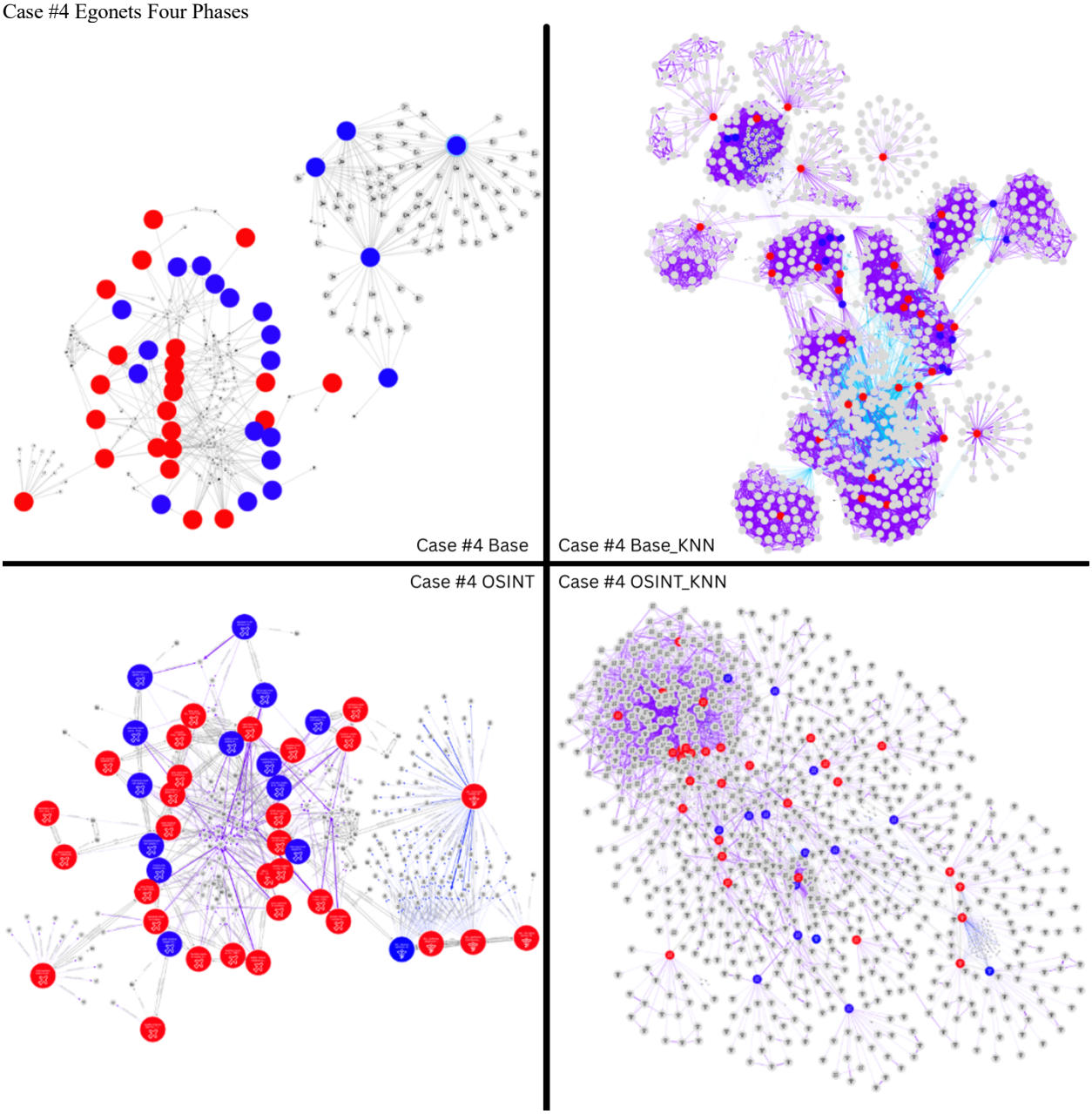


Neighborhood Statistics Case #4

Table 4-22
Case #4 Neighborhood Statistics

	Neighborhood Statistics Case #4			
	Base	Base_KNN	OSINT	OSINT_KNN
Density	0.021	0.025	0.024	0.007
Average Degree	4.34	27.22	4.86	7.39
Average Weighted Degree	339,688.68	1,578,547.49	461,004.3	208,925.89
Clustering Coefficient	0.131	0.604	0.237	0.387
Diameter	8	7	8	7
Average Path Length	3.28	3.94	3.82	4.02

Figure 4-26
Case #4 Egonets Four Phases



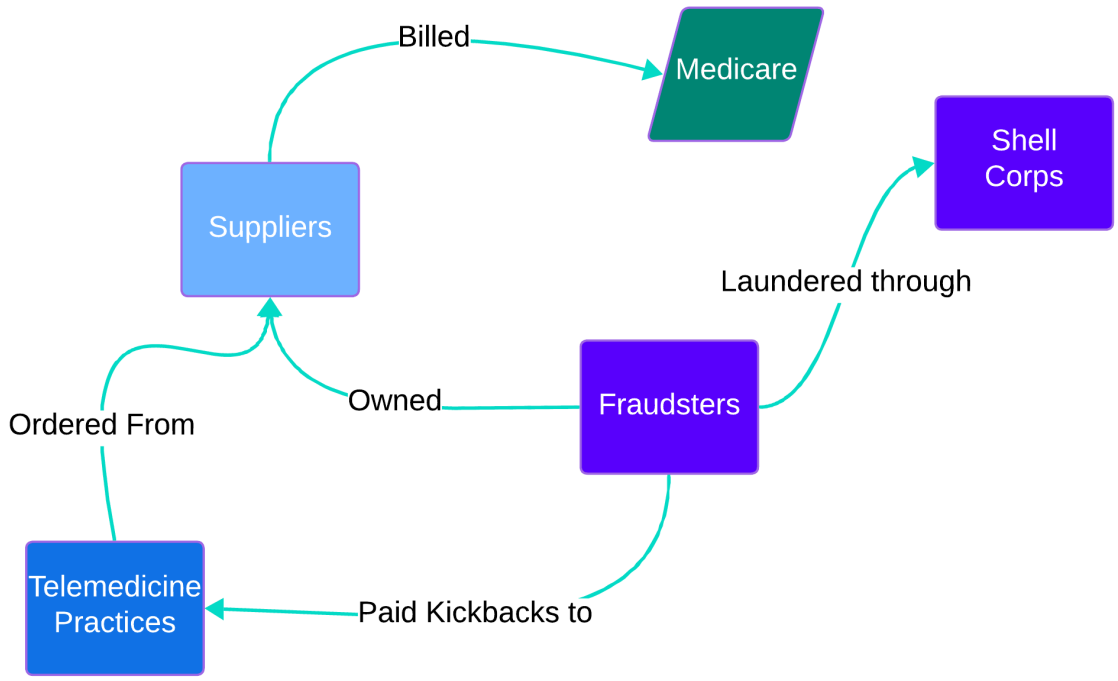
Case #5 Stchastlivtseva

Fraud Type: Kickbacks + Unnecessary Services + DME + Identity Theft
Total Loss: $17 million to Medicare. $231 million in gold, luxury vehicles, and other assets seized.
OSINT Sources: Court documents, OpenCorporates, New York Business Registry
Pre/Post Node Count: 4 8

In December 2022, Anna Stchastlivtseva was indicted as part of an ongoing Medicare fraud investigation. Stchastlivtseva owned several DME companies and allegedly paid kickbacks to purported telemedicine and marketing companies in exchange for DME orders. She and her co-conspirators then laundered the funds through several shell corps (United States vs. Stchastlivtseva).

Scheme Diagram Case #5

Figure 4-27
Case #5 Scheme Diagram

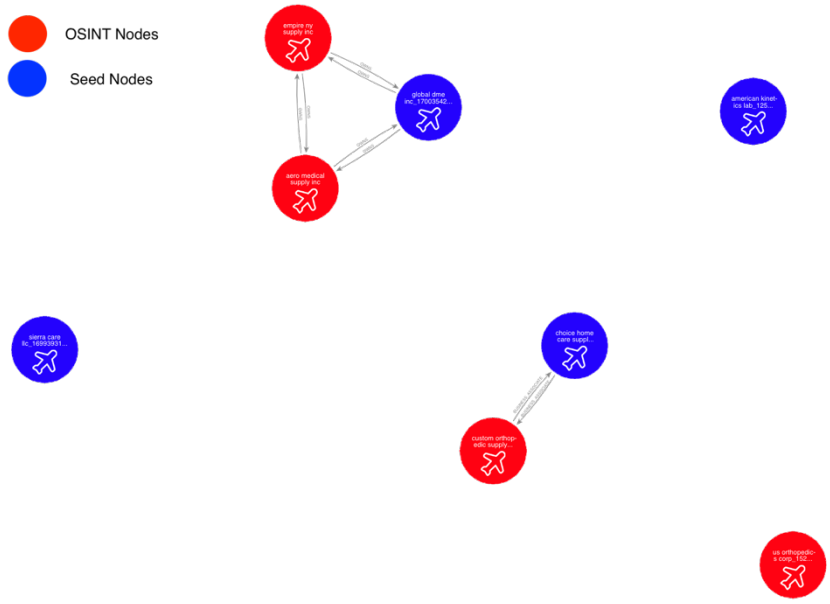


Egonet for Case #5

Table 4-23
Case #5 Egonet

Node	Type	NPI#
American Kinetics Lab	Supplier	1255761318
Sierra Care LLC	Supplier	1699393181
Global DME Inc	Supplier	1700354255
Choice Home Care Supply Inc	Supplier	1871101865
aero medical supply inc	Supplier	1821560830
empire ny supply inc	Supplier	1679196851
us orthopedics corp	Supplier	1528570033
custom orthopedic supply	Supplier	1093232811

Figure 4-28
Case #5 Supplier/Provider Nodes



Case #5 OSINT

Table 4-24
Case #5 OSINT

OSINT Name	OSINT NPI	OSINT SOURCE	SUMMARY
Aero medical supply inc	1821560830	OpenCorporates	Same registered agent as Global DME
Empire NY Supply Inc	1679196851	OpenCorporates	Same registered agent as Global DME
US Orthopedics Corp	1528570033	Indictment	Owned by defendant in a related case (Nelly Petrosyan)
Custom Orthopedic Supply	1093232811	NY Business Registry	Same address as Choice Home Case Supply. Authorized signers share same last name.

Jaccard Similarity Case #5

Figure 4-29
Case #5 Jaccard Similarity

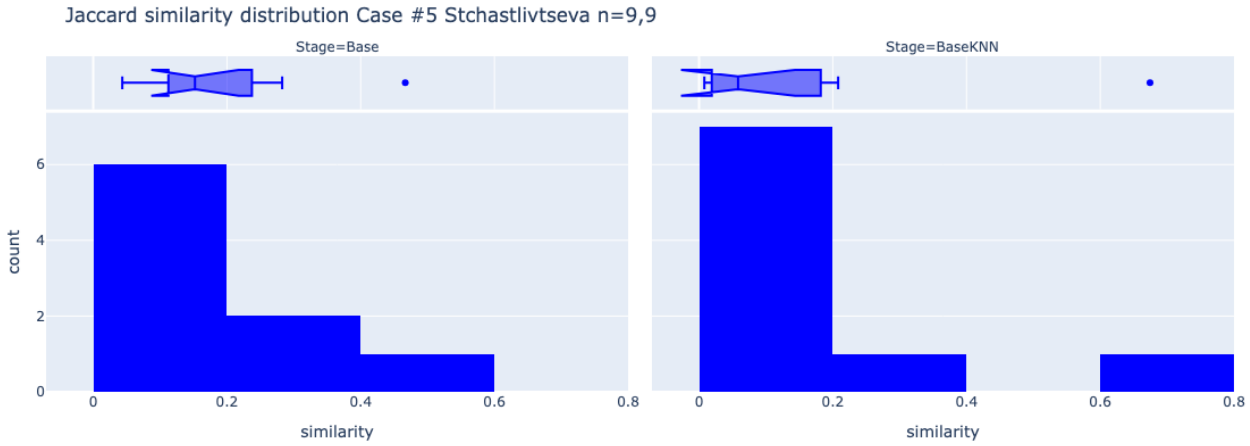
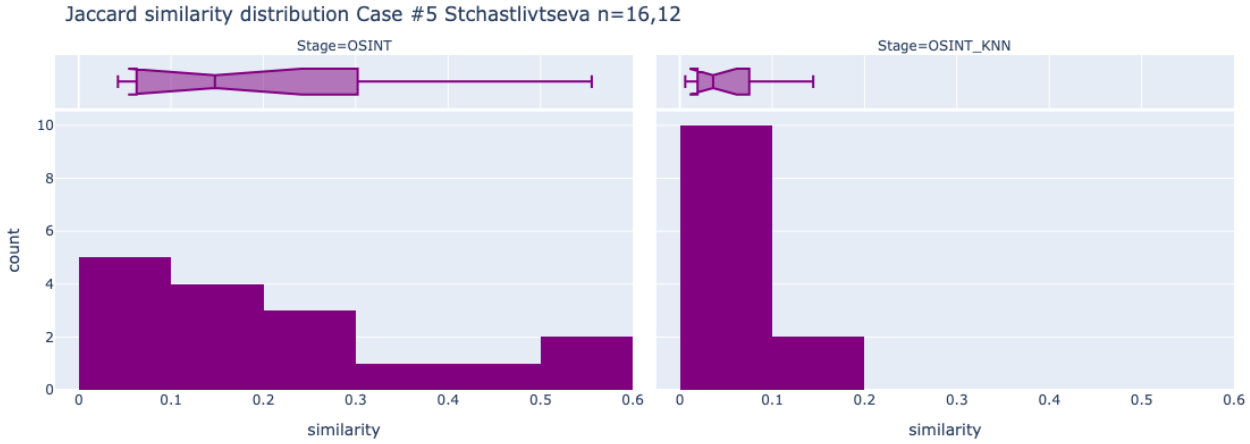


Figure 4-30
Case #5 Jaccard Similarity Post-OSINT

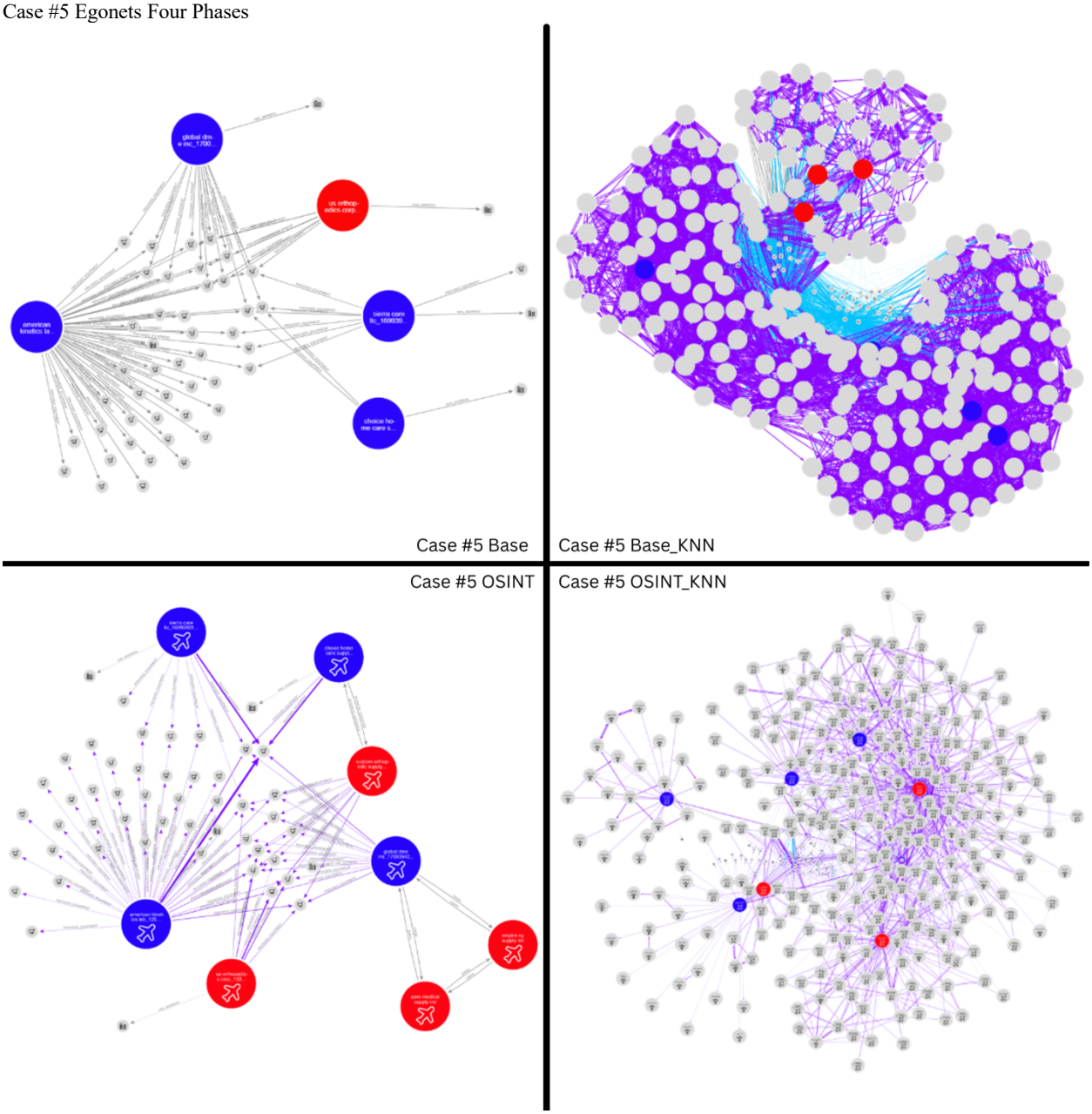


Neighborhood Statistics Case #5

Table 4-25
Case #5 Neighborhood Statistics

Neighborhood Statistics Case #5				
	Base	BaseKNN	OSINT	OSINT_KNN
Density	0.05	0.12	0.02	0.02
Average Degree	2.84	32.26	3.05	7.42
Average Weighted Degree	273,165.06	2,605,372.53	1,835,348.13	482,474.25
Clustering Coefficient	0.00	0.56	0.27	0.32
Diameter	6.00	5.00	5.00	5.00
Average Path Length	2.34	2.16	2.33	2.87

Figure 4-31
Case #5 Egonets Four Phases

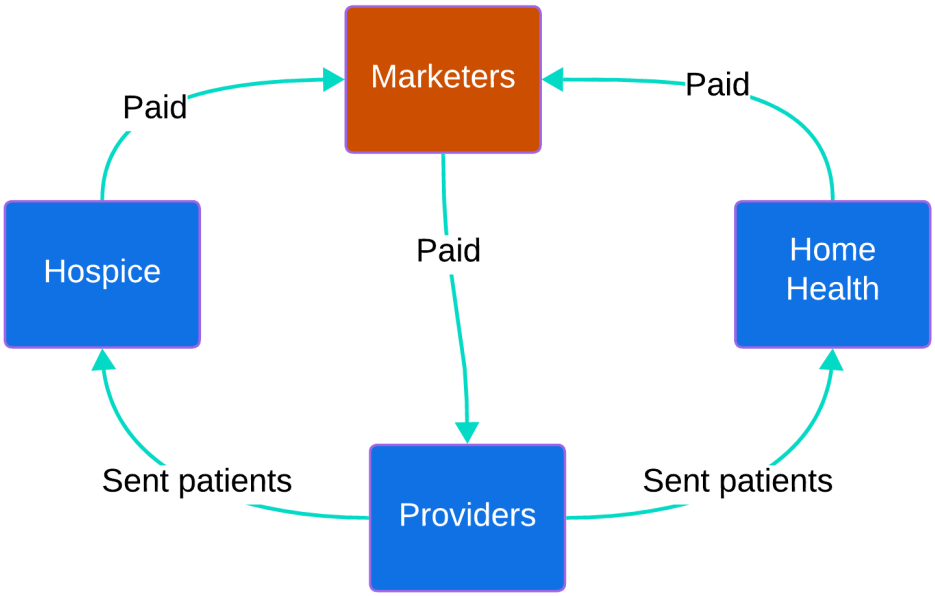


Case #6 Amity

Fraud Type: Kickbacks + Unnecessary Services
Total Loss: $300 million+, unknown # of patients who were denied curative treatment
OSINT Sources: Court documents, OpenCorporates, New York Business Registry
Pre/Post Node Count: 3 15

Scheme Diagram Case #6

Figure 4-32
Amity Fraud Scheme



Egonet for Case #6

Table 4-26
Case #5 Neighborhood Statistics

Node	Type	NPI
Bhupinder Bhandari	Provider	1417068511
Gerald Myint	Provider	1073792172
Juan Posada	Provider	1265496301
Henry Watson	Provider	1235142894

Node	Type	NPI
Arkady Massen	Provider	1306943626
Mariam Hasan	Provider	1851604847
Andre Gay	Provider	1225290752
April Mancuso	Provider	1902121429
Yelena Kabanskaya	Provider	1891006714
Kerisimasi Reynolds	Provider	1518282920
Tam Nguyen	Provider	1962520478
Scott Taylor	Provider	1548278468
Kimberly Hicks	Provider	1972661254
Rajiv Ahuja	Provider	1962460782
Marzia Mujaddide	Provider	1548508369

Case #6 OSINT

Table 4-27
Case #6 OSINT

OSINT Name	OSINT NPI	OSINT SOURCE	SUMMARY
Henry Watson	1235142894	Court Documents	Friendly with network of providers who were open to kickbacks. Introduced
Arkady Massen	1306943626	Indictment against Glennda Santos	Santos (marketer) is quoted as saying that Dr. Massen is known to participate in kickbacks
Mariam Hasan	1851604847	Court Documents	
Andre Gay	1225290752	Court Documents	Referenced as having partaken in previous kickback schemes
April Mancuso	1902121429		
Yelena Kabanskaya	1891006714		
Kerisimasi Reynolds	1518282920		
Tam Nguyen	1962520478		
Scott Taylor	1548278468	Court Documents	Dr. Taylor was introduced by Dr. Watson
Kimberly Hicks	1972661254	Court Documents	Introduced by Dr. Watson
Rajiv Ahuja	1962460782	OpenCorporates	
Marzia Mujaddide	1548508369	OpenCorporates	

Jaccard Similarity Case #6

Figure 4-33
Jaccard Similarity Case #6

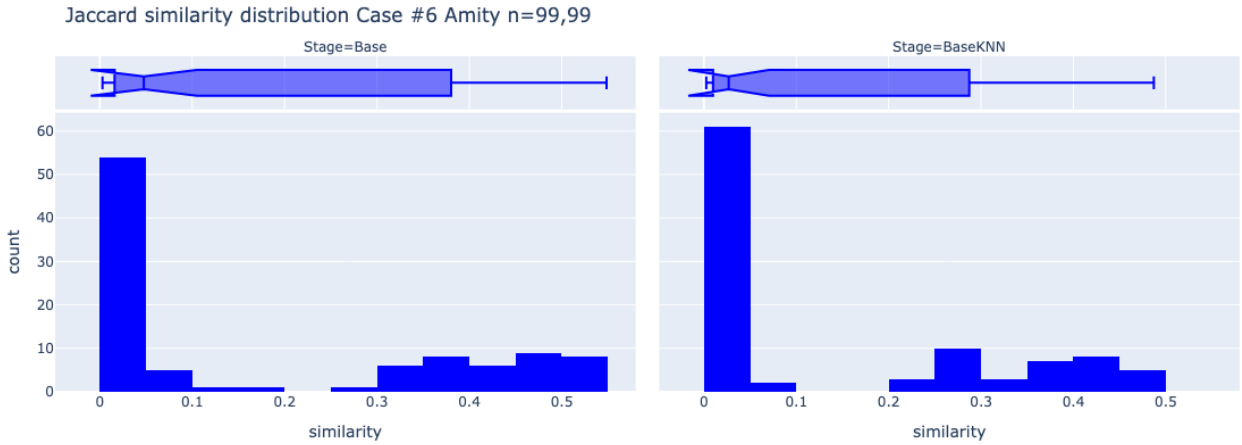
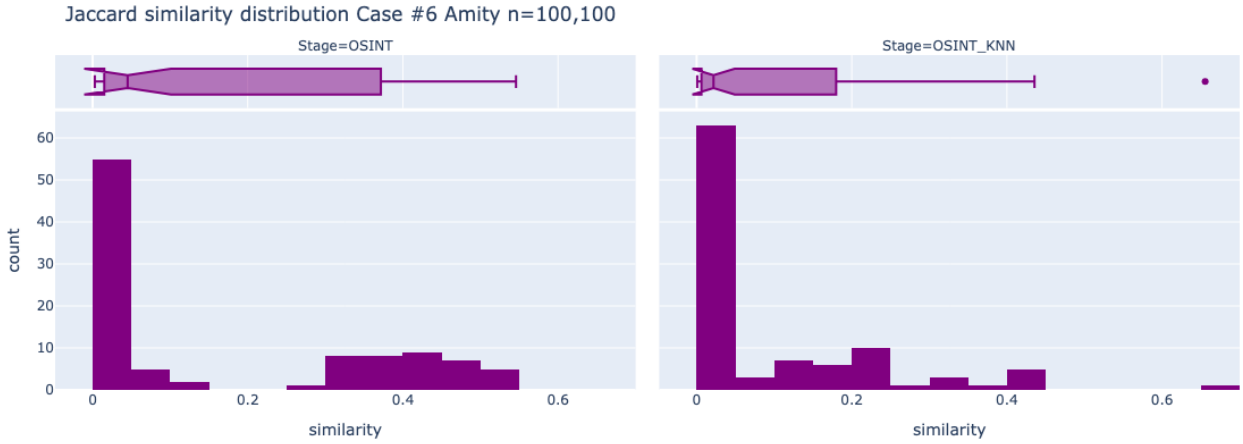


Figure 4-34
Case #6 Jaccard Similarity Post-OSINT

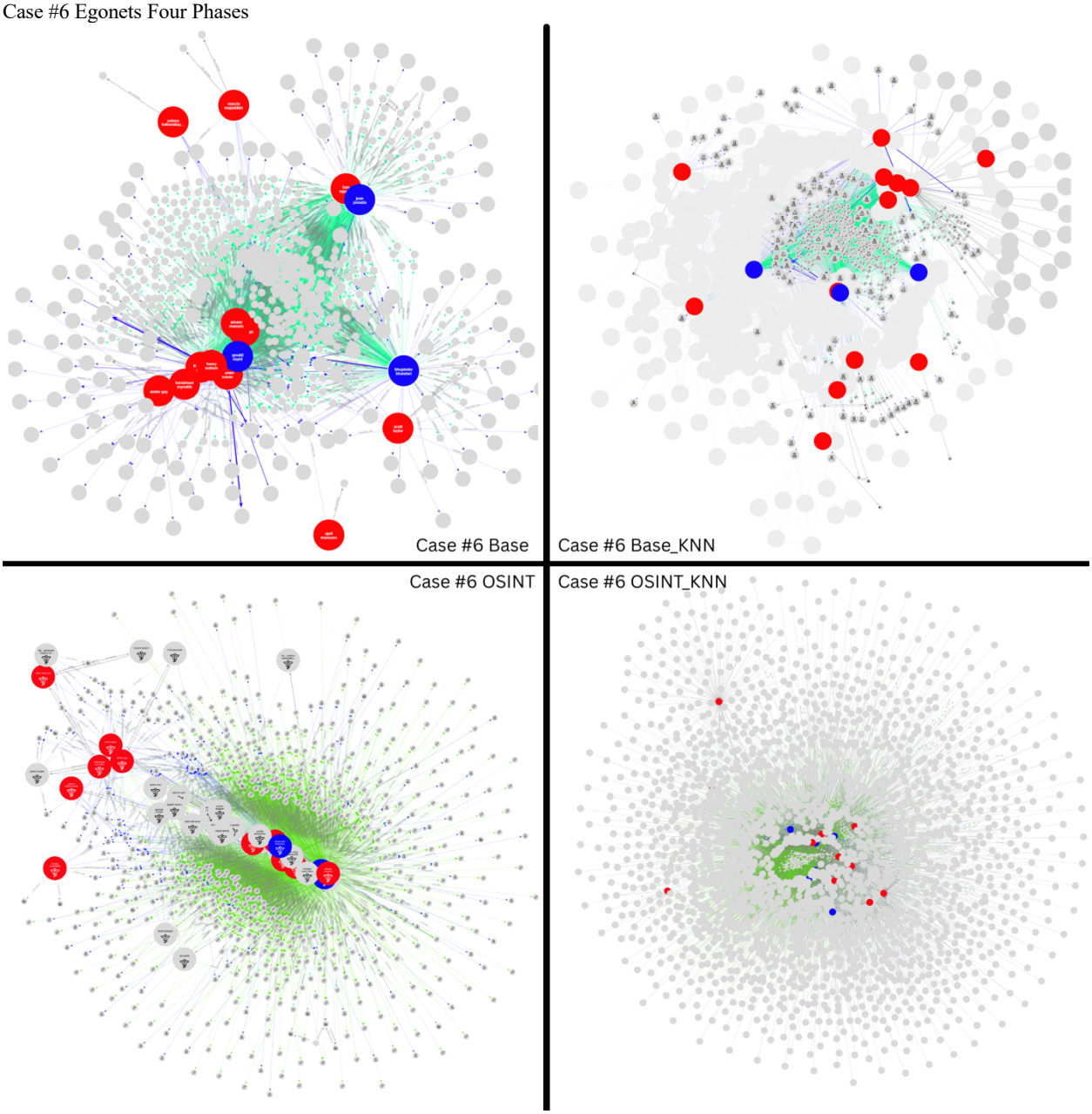


Neighborhood Statistics Case #6

Table 4-28
Case #6 Neighborhood Statistics

Neighborhood Statistics Case #6				
	Bare	BareKNN	OSINT	OSINTKNN
Density	0.011	0.032	0.017	0.012
Average Degree	6.998	42.335	10.717	35.959
Average Weighted Degree	99,072.90	315,261.36	152,854.25	146,858.78
Clustering Coefficient	0.003	0.12	0.103	0.332
Diameter	6	3	5	5
Average Path Length	2.729	2.578	2.594	2.521

Figure 4-35
Case #6 Egonets Four Phases



CHAPTER 5. DISCUSSION, IMPLICATIONS, RECOMMENDATIONS

Overview

This chapter presents a discussion of the study's results, their implications, and evidence-based recommendations for enhancing the investigation of organized Medicare fraud. The research was driven by a central problem: detecting and investigating fraud rings in high-volume, complex systems like healthcare remains deeply challenging due to their adaptive, coordinated behaviors and the limitations of siloed, single-source data. The literature review emphasized how existing detection methods often fail to account for the emergent and networked nature of fraud, particularly in systems where incomplete visibility hinders pattern recognition. Extending and applying this, the study applied a graph-based approach enriched with Open-Source Intelligence (OSINT) to uncover structural features of confirmed fraud actors in Medicare claims data. These findings reflect key principles from complexity science such as emergence, adaptation, and nonlinear interdependence, even if not explicitly invoked in the research design. Recognizing fraud as a complex system behavior offers deeper insight into why traditional detection systems struggle, and underscores the importance of multi-source, systems-level approaches that can better trace how fraudulent networks form, persist, and reconfigure.

Results

While the sample of labeled nodes was necessarily small due to the investigative time required per case, a post-hoc power analysis confirmed that the Wilcoxon tests were fully

powered (power = 1.00) to detect the observed effects. This, along with large effect sizes across all comparisons, mitigates concerns about the reliability of the results due to sample limitations.

Contribution of the Study

This study contributes to the growing body of graph-based fraud analytics by empirically testing whether augmenting Medicare claims graphs with Open-Source Intelligence (OSINT) enhances the clustering of known fraud rings. Whereas prior research has largely relied on internal datasets or domain-specific heuristics, this study introduces a scalable method for enriching public claims data with contextual information from open sources such as corporate registries, geographic indicators, and professional affiliations. This broader and more relational lens offers new insights into the formation and detection of fraud networks.

Although it may seem intuitive that adding relational data between providers increases the likelihood of their being grouped within the same community, this effect has not previously been empirically validated. Nor has prior work examined the structure of coordinated fraud across millions of provider nodes when Medicare claims are modeled as a graph. While thousands of providers may display similar billing patterns, structural context and interconnectivity offer a fundamentally different and potentially more revealing perspective on organized fraud.

By systematically testing the impact of data enrichment on community detection outcomes using the Leiden algorithm, the study adds methodological clarity to the often-overlooked role of graph construction in fraud analytics. It demonstrates that even subtle changes

in how a network is defined can materially affect which entities surface as anomalous or suspicious.

Though not explicitly framed as a complexity science study, the findings align with the view that fraud networks behave as adaptive systems exhibiting properties such as regeneration, hidden connectivity, and resilience. In doing so, this research helps bridge technical graph methods with broader systems thinking, encouraging future work that situates fraud detection within the science of complexity and adaptive networks.

Discussion and Implications

The analysis also highlights the sensitivity of community detection outcomes to network construction choices. When subtle enrichment strategies result in materially different clustering outcomes, it suggests that the success or failure of fraud analytics hinges as much on how networks are built as on the algorithms used to analyze them. This highlights the need for more transparency, rigor, and experimentation in preprocessing and feature selection stages which are areas often underreported in fraud analytics literature.

Conceptually, the study contributes to a deeper understanding of fraud as a structural and relational phenomenon, rather than merely a behavioral anomaly. Organized fraud rings are strategically embedded within networks and often display resilient, regenerative properties. The enriched graph structures in this study expose those hidden ties and overlapping affiliations that transactional data alone might obscure.

These insights also have implications for resource allocation in investigative settings. Rather than triaging individual anomalies in isolation, enriched graph analytics allow investigators to surface entire clusters for closer scrutiny potentially leading to more efficient use of limited enforcement or audit resources. In high-volume domains like healthcare, where false positives can be costly and manual investigations are time-intensive, the ability to prioritize networks rather than individuals is a significant operational advantage.

The challenge with Medicare fraud is not in detecting it, but in investigating it. All of the fraud rings analyzed in this study operated across multiple years, with new providers added over time. Some excluded providers had previous Medicare fraud convictions but were able to resume billing activity less than a decade later. This finding suggests not only investigational delays but also systemic blind spots that enable recidivism. Even when fraud patterns are known or detectable, enforcement mechanisms often fail to act quickly enough, or with enough context to fully dismantle fraudulent networks.

Finally, while the study does not explicitly operationalize complexity science, its findings resonate strongly with complexity-informed views of fraud. The emergent patterns revealed through graph enrichment suggest that fraud networks may evolve in ways that reflect adaptive, self-organizing systems– further justifying the need for analytical approaches that can capture nonlinear and multi-scalar relationships.

Limitations of the study

A major limitation of the study is that confirmation of fraud ring membership relies on conviction and official exclusion from the Medicare program. With thousands of unprosecuted and undetected cases amassing each year, the Medicare exclusionary database is not a complete representation of all fraudulent actors in the Medicare claims data set.

Another limitation of the study is that it relies on publicly available data and does not account for the possibility of falsified medical credentials, false identities, straw ownership of shell corporations, or unnamed defendants some of whom may still be working with law enforcement as part of their plea agreement. These factors create significant gaps in the understanding of the full scope of fraudulent activities, making it difficult to draw comprehensive conclusions about the effectiveness of augmenting Medicare claims with OSINT. A final limitation of this study is the reliance on historical data, which may not accurately reflect current trends in fraud schemes or the evolving tactics employed by fraudulent actors.

REFERENCES

Akerlof, G. A. (1970). The market for “lemons”: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, *84*(3), 488.

Akhgar, B., Bayerl, P. S., & Sampson, F. (Eds.). (2017). *Open source intelligence investigation: From strategy to implementation* [PDF]. Springer International Publishing.

Association of Certified Fraud Examiners. (2022). *2022 Report to the Nations*. ACFE. https://acfepublic.s3.us-west-2.amazonaws.com/2022+Report+to+the+Nations.pdf

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

Barone, M., & Coscia, M. (2018). Birds of a feather scam together: Trustworthiness homophily in a business network. *Social Networks*, *54*, 228–237.

Bauder, R., Khoshgoftaar, T. M., & Seliya, N. (2017). A survey on the state of healthcare upcoding fraud analysis and detection. *Health Services & Outcomes Research Methodology*, *17*(1), 31–55.

Browne, T. O., Abedin, M., & Chowdhury, M. J. M. (2024). A systematic review on research utilising artificial intelligence for open source intelligence (OSINT) applications. *International Journal of Information Security*, *23*(4), 2911–2938.

Chalicheemala, D., & Chalicheemala, D. (2022). What is Open-Source Intelligence and How it Can Prevent Frauds. In *papers.ssrn.com*. https://doi.org/10.2139/ssrn.4170882

Creswell, J. W., & Creswell, J. D. (2022). *Research design* (6th ed.). SAGE Publications.

Crossley, N., Bellotti, E., Edwards, G., Everett, M. G., Koskinen, J., & Tranmer, M. (2015). *Social network analysis for ego-nets: Social network analysis for actor-centred networks*.

Sage Publications. https://ebookcentral.proquest.com/lib/marymountu/reader.action?docID=5613667&ppg=37

Hairol Anuar, S. H., Abal Abas, Z., & Md Yunos, N. (2024). Identifying communities with modularity metric using louvain and leiden algorithms. *Pertanika Journal of Science & Technology*, *32*(3), 1285–1300.

Hall, R. E. (2014). Trade With Asymmetric Information. *Research Papers in Economics*, *19*, 151–160.

Hancock, J. T., Bauder, R. A., Wang, H., & Khoshgoftaar, T. M. (2023). Explainable machine learning models for Medicare fraud detection. *Journal of Big Data*, *10*(1), 154.

Huddart, S. J., & Ke, B. (2007). Information asymmetry and cross-sectional variation in insider trading. *Contemporary Accounting Research*, *24*(1), 195–232.

Isahak, M. S., Roslan, N. A. H., Abdul Tahrim, N. S. I., Zawari, S. A., Mohd Najib, W. N. A., & Lajuni, N. (2023). Factors influencing fraudulent in financial reporting using fraud triangle theory in Malaysia: A conceptual paper. *International Journal of Academic Research in Business and Social Sciences*, *13*(6). https://doi.org/10.6007/ijarbss/v13-i6/17291

Jesus, R. V., Silva, D. A. D., Torres, J. A. S., Mendonça, F. L. L. de, & de Sousa, R. T. (2023). Open source intelligence: Classification and mitigation of risks and fraud within financial institutions. *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–5.

Klein, L. S., O'Brien, T. J., & Peters, S. R. (2002). Debt vs. Equity and asymmetric information:

A review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.305401

Kramer, O. (2011). Dimensionality reduction by unsupervised K-nearest neighbor regression. *2011 10th International Conference on Machine Learning and Applications and Workshops*, *1*, 275–278.

Leskovec, J., Lang, K. J., & Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. *Proceedings of the 19th International Conference on World Wide Web*, 631–640.

Liu, X., Cheng, H.-M., & Zhang, Z.-Y. (2018). Evaluation of community detection methods. In *arXiv [cs.SI]*. arXiv. http://arxiv.org/abs/1807.01130

Masihullah, S., Negi, M., Matthew, J., & Sathyanarayana, J. (2022). Identifying Fraud Rings Using Domain Aware Weighted Community Detection. *Machine Learning and Knowledge Extraction: 6th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2022, Vienna, Austria, August 23–26, 2022, Proceedings*, 150–167.

Meyers, T. J. (2017). Examining the network components of a Medicare fraud scheme: the Mirzoyan-Terdjanian organization. *Crime, Law, and Social Change*, *68*(1), 251–279.

O'Malley, A. J., Bubolz, T. A., & Skinner, J. S. (2023). The diffusion of health care fraud: A bipartite network analysis. *Social Science & Medicine*, *327*, 115927.

Ouellet, M., Maimon, D., Wu, Y., Howell, C. J., Abay, D., Bharthepudi, H., Bondalapati, A., Chen, X., Crumpler, M., Darbha, S., Divyakolu, S., Gadde, O., Harrison, T., Kalluri, M., Kambhampati, S., Kodali, M., Malapati, V. B., Mueller, R., Rai, K., … Stubler, N. (2022, June 30). *Open source intelligence in online stolen data markets: Assessment of network*

disruption strategies. PubPub. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=GqggT9MAAAAJ&sortby=pubdate&citation_for_view=GqggT9MAAAAJ:ZHo1McVdvXMC

Pastor-Galindo, J., Nespoli, P., Gomez Marmol, F., & Martinez Perez, G. (2020). The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 10282–10304.

Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, *133*, 113303.

Quach, N. E., Yang, K., Chen, R., Tu, J., Xu, M., Tu, X. M., & Zhang, X. (2022). Post-hoc power analysis: a conceptually valid approach for power based on observed study data. *General Psychiatry*, *35*(4), e100764.

Ran, X., Meara, E., Morden, N. E., Moen, E. L., Rockmore, D. N., & O'Malley, A. J. (2024). Estimating the impact of physician risky-prescribing on the network structure underlying physician shared-patient relationships. *Applied Network Science*, *9*(1), 63.

Reis, E. F. dos, Teytelboym, A., ElBahraw, A., De Loizaga, I., & Baronchelli, A. (2023). Identifying key players in dark web marketplaces. In *arXiv [physics.soc-ph]*. arXiv. http://arxiv.org/abs/2306.09485

Sarvari, H., Abozinadah, E., Mbaziira, A., & Mccoy, D. (2014). Constructing and Analyzing Criminal Networks. *2014 IEEE Security and Privacy Workshops*, 84–91.

Schneider, K. C., & Kerlinger, F. N. (1979). Behavioral research: A conceptual approach. *JMR, Journal of Marketing Research*, *16*(4), 599.

Shekhar, S., Leder-Luis, J., & Akoglu, L. (2023). *Unsupervised machine learning for explainable health care fraud detection* (No. 30946). National Bureau of Economic Research. https://www.nber.org/papers/w30946

Shishkov, P., Kanaeva, M., & Lozhechko, A. (2022). Analysis of the practical use of information asymmetry in financial markets. *Russian Journal of Resources Conservation and Recycling*, *9*(4). https://doi.org/10.15862/41ecor422

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233.

Travieso, G., Benatti, A., & Costa, L. da F. (2024). An analytical approach to the Jaccard similarity index. In *arXiv [physics.data-an]*. arXiv. http://arxiv.org/abs/2410.16436

Tri Wijaya, J. R., & Herwiyanti, E. (2023). A study of information asymmetry in financial research. *The Indonesian Accounting Review*, *13*(1), 79–89.

Wang, X., Xiangfeng, L., Wang, X., & Yu, H. (2024). Homophilic and heterophilic-aware sparse graph transformer for financial fraud detection. *2024 International Joint Conference on Neural Networks (IJCNN)*, *29*, 1–8.

Yaacob, M. H., Thing, N. S., & Alias, N. (2024). Bridging the gap between information asymmetry and IR4.0: A systematic literature review. In *Contemporary Issues in Finance, Investment and Banking in Malaysia* (pp. 1–13). Springer Nature Singapore.

Yadav, A., Kumar, A., & Singh, V. (2023). Open-source intelligence: a comprehensive review of the current state, applications and future perspectives in cyber security. *Artificial Intelligence Review*, 1–32.

Zamost, S., & Brewer, C. (2023, March 9). *Inside the mind of criminals: How to brazenly steal*

$100 billion from Medicare and Medicaid. CNBC. https://www.cnbc.com/2023/03/09/how-medicare-and-medicaid-fraud-became-a-100b-problem-for-the-us.html

Zhang, J. (2024). A literature review on the theory of asymmetric information. *Advances in Economics, Management and Political Sciences*, *124*(1), 183–189.

References

Abdullahi, R., & Mansor, N. (2018). Fraud prevention initiatives in the Nigerian public sector: Understanding the relationship of fraud incidences and the elements of fraud triangle theory. *Journal of Financial Crime*, 25, 527-544. doi:10.1108/JFC- 02-2015-0008

Abulencia, J. (2021). The cost of cybercrime in the US healthcare sector. *Computer Fraud & Security*, *2021*(11), 8–13.

Baesens, B., Van Vlasselaer, V., & Verbeke, W. (2015). *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection (Wiley and SAS Business Series)* (1st ed.). Wiley.

Bevan, R. (2015). *The Changemaking Checklists: A Toolkit for Planning, Leading, and Sustaining Change*. ChangeStart Press.

Bossler, A. M., & Berenblum, T. (2019). Introduction: new directions in cybercrime research. *Journal of Crime and Justice*, *42*(5), 495–499.

Bulakh, V. (2017). *Online fraud economy: Characterization and defense* (M. Gupta (ed.)) [Indiana University].

Burke, A., van Stel, A., & Thurik, R. (2010). Blue Ocean vs. Five Forces. Harvard Business Review. https://hbr.org/2010/05/blue-ocean-vs-five-forces

Burrell, D. N., Nobles, C., Cusak, A., & Omar, M. (2022). Cybercrime and the Nature of Insider Threat Complexities in Healthcare and Biotechnology Engineering Organizations. *Journal of Crime and Criminal Behavior.*

Compin, F. (2016). Do financial criminals commit perfect crimes? *Journal of Financial Crime*, 23, 624-636. doi:10.1108/JFC-03/2015/0018

Cybersecurity is Patient Safety. (2022). US Senate Intelligence Committee. https://www.warner.senate.gov/public/_cache/files/f/5/f5020e27-d20f-49d1-b8f0-bac298f5da0b/0320658680B8F1D29C9A94895044DA31.cips-report.pdf

Daramola, O. E., Abu, J. M., Daramola, L. O., & Akande, T. M. (2019). Medical Identity Fraud in Health Insurance Schemes: Creating Awareness in Nigeria. Asian Journal of Case Reports in Medicine and Health, 2 (1) , 1-6.

Davis, J. (2019a, January 31). Phishing hack breaches PHI of 23,000 Colorado patients for 3 months. Retrieved from https://healthitsecurity.com/news/phishing-hack-breaches-phi-of-23000-colorado-patients-for-3-months

Ebrahimi, M., Surdeanu, M., Samtani, S., & Chen, H. (2018). Detecting Cyber Threats in Non-English Dark Net Markets: A Cross-Lingual Transfer Learning Approach. *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 85–90.

Federal Trade Commission. Taking Charge: What to do if your identity is stolen. Federal Trade Commission Website. Available at https://www.consumer.ftc.gov/articles/pdf-0009-taking-charge.pdf

Furlan, S., & Bajec, M. (2008). Holistic approach to fraud management in health insurance. *Journal of Information and Organizational Sciences*, 32(2), 99-114.

Genes, R. (2016). Code cyber: Preventing breaches at hospitals and health care practices. *Journal of Health Care Compliance*, 18(3), 13-18.

Harwell, D. (2023, April 15). The military loved Discord for Gen Z recruiting. Then the leaks began. *The Washington Post*. https://www.washingtonpost.com/technology/2023/04/15/discord-military-recruitment-

pentagon-document-leaks/

Hibbard, D., Hibbard, M., & Stockman, J. (2006). *The Canoe Theory: A Business Success Strategy for Leaders and Associates* (0 ed.). iUniverse, Inc.

Jiang, R. (2022). Exploring Employees' Computer Fraud Behaviors using the Fraud Triangle Theory. *Pacific Asia Journal of the Association for Information Systems*, *14*(4), 4.

Johnson, C. E. (2020). *Meeting the Ethical Challenges of Leadership: Casting Light or Shadow* (7th ed.). SAGE Publications, Inc.

Katz, B. (2018). Connecting care for patients: Interdisciplinary care transition and collaboration. Guilford, CT: Jones & Bartlett Learning.

Kennedy, J. P. (2017). Functional redundancy as a response to employee theft within small businesses. *Security Journal; London*, *30*(1), 162–183.

Korgaonkar, P., Becerra, E. P., Mangleburg, T., & Bilgihan, A. (2021). Retail employee theft: When retail security alone is not enough. *Psychology & Marketing*, *38*(5), 721–734.

Labong, R. C. (2019). Identity Theft Protection Strategies: A Literature Review. *Journal of Academic Research*, *4*(2), 1–12.

Liu, Y., Lin, F. Y., Ahmad-Post, Z., Ebrahimi, M., Zhang, N., Hu, J. L., Xin, J., Li, W., & Chen, H. (2020). Identifying, Collecting, and Monitoring Personally Identifiable Information: From the Dark Web to the Surface Web. *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 1–6.

Lokanan, M. E. (2018). Informing the fraud triangle: Insights from differential association theory. *Journal of Theoretical Accounting Research*, 14(1), 55-98.

Office of Information Security. (2023, March 9). *Data Exfiltration Trends in Healthcare*. Data Exfiltration Trends in Healthcare. https://www.hhs.gov/sites/default/files/data-exfiltration-in-healthcare-tlpclear.pdf

Association of Certified Fraud Examiners. 2022. "2022 Report to the Nations." ACFE. https://acfepublic.s3.us-west-2.amazonaws.com/2022+Report+to+the+Nations.pdf.

Cybersecurity Ventures. 2022. "2022 Official Cybercrime Report." Cybersecurity Ventures. https://s3.ca-central-1.amazonaws.com/esentire-dot-com-assets/assets/resourcefiles/2022-Official-Cybercrime-Report.pdf.

Gehl, Robert W. 2016. "Power/Freedom on the Dark Web: A Digital Ethnography of the Dark Web Social Network." *New Media & Society* 18 (7): 1219–35.

Gonzalez, Juan, Ignacio Garijo, and Alfonso Sanchez. 2020. "Organ Trafficking and Migration: A Bibliometric Analysis of an Untold Story." *International Journal of Environmental Research and Public Health* 17 (9). https://doi.org/10.3390/ijerph17093204.

Grover, Prince, Zheng Li, Jianbo Liu, Jakub Zablocki, Hao Zhou, Julia Xu, and Anqi Cheng. 2022. "FDB: Fraud Dataset Benchmark." *ArXiv [Cs.LG]*. arXiv. https://github.com/amazon-research/fraud-dataset-benchmark.

International Labor Organization. 2022. "Global Estimates of Modern Slavery." ILO. https://www.ilo.org/wcmsp5/groups/public/---ed_norm/---ipec/documents/publication/wcms_854733.pdf.

Kurshan, Eren, Hongda Shen, and Haojie Yu. 2020. "Financial Crime & Fraud Detection Using Graph Computing: Application Considerations & Outlook." In *2020 Second International Conference on Transdisciplinary AI (TransAI)*, 125–30. ieeexplore.ieee.org.

Masihullah, Shaik, Meghana Negi, Jose Matthew, and Jairaj Sathyanarayana. 2022. “Identifying Fraud Rings Using Domain Aware Weighted Community Detection.” In *Machine Learning and Knowledge Extraction*, 150–67. Springer International Publishing.

Omair, Badr, and Ahmad Alturki. 2020. “Taxonomy of Fraud Detection Metrics for Business Processes.” *IEEE Access* 8: 71364–77.

Paschal, Uchenna, Wilson Nwankwo, Florence U. Masajuwa, Simon Imoisi, and Paschal Uchenna Chinedu. 2021. “Cybercrime Detection and Prevention Efforts in the Last Decade: An Overview of the Possibilities of Machine Learning Models.” *Review Of.* https://doi.org/10.48047/rigeo.11.07.92.

Poese, Ingmar, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet, and Bamba Gueye. 2011. “IP Geolocation Databases: Unreliable?” *SIGCOMM Comput. Commun. Rev.* 41 (2): 53–56.

Weinberg, Zachary, Shinyoung Cho, Nicolas Christin, Vyas Sekar, and Phillipa Gill. 2018. “How to Catch When Proxies Lie: Verifying the Physical Locations of Network Proxies with Active Geolocation.” In *Proceedings of the Internet Measurement Conference 2018*, 203–17. IMC ’18. New York, NY, USA: Association for Computing Machinery.

Weinglass, Simona. 2016. “The Wolves of Tel Aviv: Israel’s Vast, Amoral Binary Options Scam Exposed.” *The Times of Israel*, March 23, 2016. https://www.timesofisrael.com/the-wolves-of-tel-aviv-israels-vast-amoral-binary-options-scam-exposed/.

Wu, Ling, Qiong Peng, and Michael Lembke. 2023. “Research Trends in Cybercrime and Cybersecurity: A Review Research Trends in Cybercrime and Cybersecurity: A Review.” *International Journal of Cybersecurity Intelligence and Cybercrime* 6 (1): 5–28.

Akerlof, G. A. (1970). 4. The market for “lemons”: quality uncertainty and the market mechanism. *Market Failure or Success*. https://www.elgaronline.com/downloadpdf/edcollbook/1843760258.pdf#page=82

Akhgar, B., Bayerl, P. S., & Sampson, F. (Eds.). (2017). *Open source intelligence investigation: From strategy to implementation* (1st ed.) [PDF]. Springer International Publishing.

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

Barone, M., & Coscia, M. (2018). Birds of a feather scam together: Trustworthiness homophily in a business network. *Social Networks*, *54*, 228–237.

Bauder, R. A., & Khoshgoftaar, T. M. (2018). The detection of medicare fraud using machine learning methods with excluded provider labels. *The Thirty-First International Flairs Conference*. https://cdn.aaai.org/ocs/17617/17617-77660-1-PB.pdf

Browne, T. O., Abedin, M., & Chowdhury, M. J. M. (2024). A systematic review on research utilising artificial intelligence for open source intelligence (OSINT) applications. *International Journal of Information Security*, *23*(4), 2911–2938.

Chalicheemala, D., & Chalicheemala, D. (2022). What is Open-Source Intelligence and How it Can Prevent Frauds. In *papers.ssrn.com*. https://doi.org/10.2139/ssrn.4170882

Chen, H., Sultan, S. F., Tian, Y., Chen, M., & Skiena, S. (2019). Fast and accurate network embeddings via very sparse random projection. In *arXiv [cs.SI]*. arXiv. http://arxiv.org/abs/1908.11512

Creswell, J. W., & Creswell, J. D. (2022). *Research design* (6th ed.). SAGE Publications.

Crossley, N., Bellotti, E., Edwards, G., Everett, M. G., Koskinen, J., & Tranmer, M. (2015).

Social network analysis for ego-nets: Social network analysis for actor-centred networks. Sage Publications. https://ebookcentral.proquest.com/lib/marymountu/reader.action?docID=5613667&ppg=37

Federal Bureau of Investigation. (2019, April 9). *Billion-Dollar Bust*. https://www.fbi.gov/news/stories/billion-dollar-medicare-fraud-bust-040919

GAO. (2024). *Medicare and medicaid additional actions needed to enhance program integrity and save billions*. United States Government Accountability Office. https://www.gao.gov/assets/gao-24-107487.pdf

Grover, P., Li, Z., Liu, J., Zablocki, J., Zhou, H., Xu, J., & Cheng, A. (2022). FDB: Fraud Dataset Benchmark. In *arXiv [cs.LG]*. arXiv. https://github.com/amazon-research/fraud-dataset-benchmark

Hairol Anuar, S. H., Abal Abas, Z., & Md Yunos, N. (2024). Identifying communities with modularity metric using louvain and leiden algorithms. *Pertanika Journal of Science & Technology*, *32*(3), 1285–1300.

Hall, R. E. (2014). Trade With Asymmetric Information. *Research Papers in Economics*, *19*, 151–160.

Huddart, S. J., & Ke, B. (2007). Information asymmetry and cross-sectional variation in insider trading. *Contemporary Accounting Research*, *24*(1), 195–232.

Isahak, M. S., Roslan, N. A. H., Abdul Tahrim, N. S. I., Zawari, S. A., Mohd Najib, W. N. A., & Lajuni, N. (2023). Factors influencing fraudulent in financial reporting using fraud triangle theory in Malaysia: A conceptual paper. *International Journal of Academic*

Research in Business and Social Sciences, *13*(6). https://doi.org/10.6007/ijarbss/v13-i6/17291

Jesus, R. V., Silva, D. A. D., Torres, J. A. S., Mendonça, F. L. L. de, & de Sousa, R. T. (2023). Open source intelligence: Classification and mitigation of risks and fraud within financial institutions. *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–5.

Johnson, T. L., & So, E. C. (2018). A simple multimarket measure of information asymmetry. *Management Science*, *64*(3), 1055–1080.

Klein, L. S., O'Brien, T. J., & Peters, S. R. (2002). Debt vs. Equity and asymmetric information: A review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.305401

Kramer, O. (2011). Dimensionality reduction by unsupervised K-nearest neighbor regression. *2011 10th International Conference on Machine Learning and Applications and Workshops*, *1*, 275–278.

Kumar, M. S., Srivastava, V., Behera, B. B., Savariapitchai, M., Sahu, S., Mahajan, R., & George, A. S. (2024). IoE and AI in real-time customer behavior analysis. In *Advances in Computational Intelligence and Robotics* (pp. 241–256). IGI Global.

Kurshan, E., & Shen, H. (2020). Graph Computing for Financial Crime and Fraud Detection: Trends, Challenges and Outlook. *International Journal of Semantic Computing*, *14*(04), 565–589.

Leskovec, J., Lang, K. J., & Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. *Proceedings of the 19th International Conference on World Wide Web*, 631–640.

Lin, H., Zhan, Y., Zhao, Z., Chen, Y., & Dong, C. (2021). Overlapping community detection based on attribute augmented graph. *Entropy (Basel, Switzerland)*, *23*(6), 680.

Liu, X., Cheng, H.-M., & Zhang, Z.-Y. (2018). Evaluation of community detection methods. In *arXiv [cs.SI]*. arXiv. http://arxiv.org/abs/1807.01130

McNicholas, T., & Randolph, W. S. (2024, February 15). *Brooklyn business connected to multimillion dollar suspected Medicare fraud ring*. CBS New York. https://www.cbsnews.com/newyork/news/g-and-i-ortho-supply-gravesend-catheters-medicare/

Memon, M. A., Thurasamy, R., Ting, H., & Cheah, J.-H. (2024). Purposive sampling: A review and guidelines for quantitative research. *Journal of Applied Structural Equation Modeling*, *9*(1), 1–23.

Meyers, T. J. (2017). Examining the network components of a Medicare fraud scheme: the Mirzoyan-Terdjanian organization. *Crime, Law, and Social Change*, *68*(1), 251–279.

O'Malley, A. J., Bubolz, T. A., & Skinner, J. S. (2023). The diffusion of health care fraud: A bipartite network analysis. *Social Science & Medicine*, *327*, 115927.

Ouellet, M., Maimon, D., Wu, Y., Howell, C. J., Abay, D., Bharthepudi, H., Bondalapati, A., Chen, X., Crumpler, M., Darbha, S., Divyakolu, S., Gadde, O., Harrison, T., Kalluri, M., Kambhampati, S., Kodali, M., Malapati, V. B., Mueller, R., Rai, K., … Stubler, N. (2022, June 30). *Open source intelligence in online stolen data markets: Assessment of network disruption strategies*. PubPub. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=GqggT9MAAAAJ&sortby=pubdate&citation_for_view=GqggT9MAAAAJ:ZHo1McVdvXMC

Pastor-Galindo, J., Nespoli, P., Gomez Marmol, F., & Martinez Perez, G. (2020). The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 10282–10304.

Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, *133*, 113303.

Ran, X., Meara, E., Morden, N. E., Moen, E. L., Rockmore, D. N., & O'Malley, A. J. (2024). Estimating the impact of physician risky-prescribing on the network structure underlying physician shared-patient relationships. *Applied Network Science*, *9*(1), 63.

Reis, E. F. dos, Teytelboym, A., ElBahraw, A., De Loizaga, I., & Baronchelli, A. (2023). Identifying key players in dark web marketplaces. In *arXiv [physics.soc-ph]*. arXiv. http://arxiv.org/abs/2306.09485

Sarvari, H., Abozinadah, E., Mbaziira, A., & Mccoy, D. (2014). Constructing and Analyzing Criminal Networks. *2014 IEEE Security and Privacy Workshops*, 84–91.

Schneider, K. C., & Kerlinger, F. N. (1979). Behavioral research: A conceptual approach. *JMR, Journal of Marketing Research*, *16*(4), 599.

Shekhar, S., Leder-Luis, J., & Akoglu, L. (2023). *Unsupervised machine learning for explainable health care fraud detection* (No. 30946). National Bureau of Economic Research. https://www.nber.org/papers/w30946

Shishkov, P., Kanaeva, M., & Lozhechko, A. (2022). Analysis of the practical use of information asymmetry in financial markets. *Russian Journal of Resources Conservation and Recycling*, *9*(4). https://doi.org/10.15862/41ecor422

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233.

Travieso, G., Benatti, A., & Costa, L. da F. (2024). An analytical approach to the Jaccard similarity index. In *arXiv [physics.data-an]*. arXiv. http://arxiv.org/abs/2410.16436

Tri Wijaya, J. R., & Herwiyanti, E. (2023). A study of information asymmetry in financial research. *The Indonesian Accounting Review*, *13*(1), 79–89.

U.S. Department of Justice. (2024). *Florida Man Sentenced to 10 Years in Prison and Ordered to Pay More Than $97 Million in Restitution for Participation in Multiple Health Care Fraud and Kickback Schemes*. https://www.justice.gov/usao-wdpa/pr/florida-man-sentenced-10-years-prison-and-ordered-pay-more-97-million-restitution

Wang, X., Xiangfeng, L., Wang, X., & Yu, H. (2024). Homophilic and heterophilic-aware sparse graph transformer for financial fraud detection. *2024 International Joint Conference on Neural Networks (IJCNN)*, *29*, 1–8.

Wilson, R. (2023). *South Florida is "ground zero" for healthcare fraud*. https://www.beckershospitalreview.com/legal-regulatory-issues/south-florida-is-ground-zero-for-healthcare-fraud.html

Yaacob, M. H., Thing, N. S., & Alias, N. (2024). Bridging the gap between information asymmetry and IR4.0: A systematic literature review. In *Contemporary Issues in Finance, Investment and Banking in Malaysia* (pp. 1–13). Springer Nature Singapore.

Yadav, A., Kumar, A., & Singh, V. (2023). Open-source intelligence: a comprehensive review of the current state, applications and future perspectives in cyber security. *Artificial Intelligence Review*, 1–32.

Zhang, J. (2024). A literature review on the theory of asymmetric information. *Advances in Economics, Management and Political Sciences*, *124*(1), 183–189.

APPENDIX

Wilcoxon Monte Carlo Simulation

```
import pandas as pd
import numpy as np
from scipy.stats import wilcoxon
import matplotlib.pyplot as plt

# === Step 1: Sample F1 scores from Base and OSINT ===

# Assume each case has multiple resolution values.
# We'll average across those to get one F1 score per case per method.

# Filter just the F1 rows for Base and OSINT
f1_base = df[(df['period'] == 'Base') &
df['case'].str.endswith('_F1')]
f1_osint = df[(df['period'] == 'OSINT') &
df['case'].str.endswith('_F1')]

# Extract case names
case_names = f1_base['case'].str.replace('_F1', '')

# Compute per-case means across resolution columns
resolution_cols = [col for col in df.columns if
col.startswith('Leiden_')]

f1_base_mean = f1_base[resolution_cols].mean(axis=1).values
f1_osint_mean = f1_osint[resolution_cols].mean(axis=1).values

# === Step 2: Paired Wilcoxon Test ===
stat, p = wilcoxon(f1_base_mean, f1_osint_mean)
print(f"Wilcoxon signed-rank test: statistic={stat:.3f}, p={p:.4f}")

# === Optional: Effect Size (Cohen's d for paired data) ===
mean_diff = np.mean(f1_osint_mean - f1_base_mean)
std_diff = np.std(f1_osint_mean - f1_base_mean, ddof=1)
cohen_d = mean_diff / std_diff
```

```
print(f"Cohen's d (paired): {cohen_d:.3f}")

# === Step 3: Power Simulation ===
import statsmodels.stats.power as smp

def simulate_wilcoxon_power(effect_size, std_dev, n, alpha=0.05,
sims=1000):
    significant = 0
    for _ in range(sims):
        base = np.random.normal(loc=0, scale=std_dev, size=n)
        osint = base + effect_size
        stat, p = wilcoxon(base, osint)
        if p < alpha:
            significant += 1
    return significant / sims

# Run simulation to estimate needed cases for 80% power
for n in range(5, 21):
    power = simulate_wilcoxon_power(effect_size=mean_diff,
std_dev=std_diff, n=n)
    print(f"n={n}, simulated power={power:.2f}")
```

Resolution vs Modularity

The resolution parameter (γ) in the Leiden algorithm allows users to control the granularity of detected communities. Higher resolutions yield smaller, more detailed partitions, while lower resolutions produce larger, more aggregated communities. When the resolution parameter is too low, community detection may fail to capture more granular structures in large graphs.