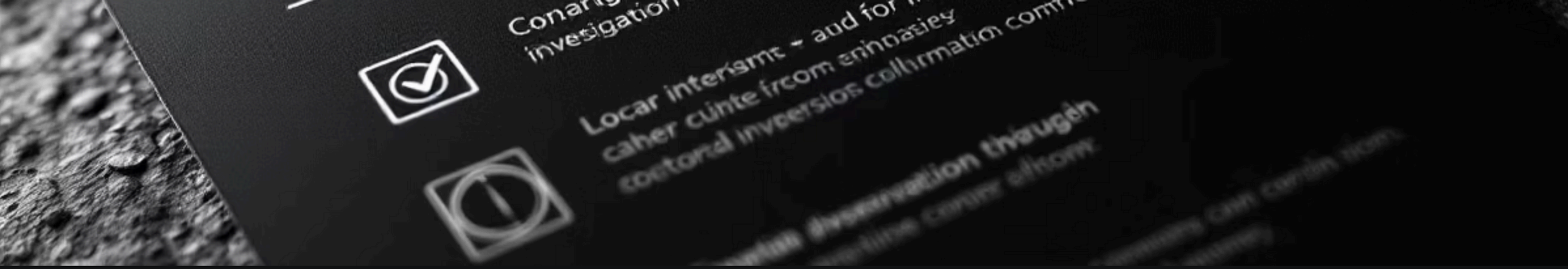
Web Scraping and OSINT Tools for Crime Investigation: Detecting Suspicious Seller Networks

Presented By:
T.U Adarsh-2023BCY0039.



Objectives

The primary objective of this project is to demonstrate how web scraping techniques can be applied to support crime investigations by identifying patterns of potentially stolen goods on online marketplaces. A privately hosted test website containing synthetic product listings will be created to simulate a real marketplace environment. Using Python tools such as Requests, BeautifulSoup, or Selenium, data including product titles, descriptions, prices, seller details, contact information, image URLs, and timestamps will be collected and analyzed. The goal is to detect repeated indicators of suspicious activity — such as reused images, identical contact details, and repeated phrases — and visualize the relationships between sellers and listings through a network map. This will highlight clusters of interconnected or potentially illicit behavior. The project also aims to document the methodology, findings, and ethical considerations, confirming that all data used is synthetic and hosted privately for academic and demonstration purposes.

Used Tools and modules

- Web Scraping (Python + BeautifulSoup)
  - Purpose: Extract product listings automatically from the synthetic marketplace website.
- Web Scraping (OctoParse)
  - Octoparse is a no-coding solution for web scraping to turn pages into structured data.
- Data Flagging & Cleaning (Pandas & OpenRefine)
  - Purpose: Clean, standardize, and flag suspicious attributes like reused contacts, images, and low prices.
- Network Visualization (Gephi & Plotly)
  - Purpose: Visualize relationships between sellers, contacts, and reused images to identify hidden clusters.
- Unsupervised Machine Learning (Algorithm-Isolation Forest)

Purpose: Predict and classify suspicious listings based on risk features for automated detection.

- OSINT Tool – Maltego
  - Purpose: Enrich investigation by mapping connections from emails, domains, and online profiles.
- OSINT Tool – Numverify
  - Purpose: Validate phone numbers to check authenticity and detect disposable or fake numbers.

Web Scraping:

- Built using Python, leveraging Requests and BeautifulSoup libraries for efficient data extraction.
- Scrapes synthetic product listings from a locally hosted website (ethical, no real data).
- Extracts key fields: title, description, price, seller handle, contact info, image URL, timestamp.
- Identifies repeated indicators such as reused images, duplicate contacts, and identical descriptions
- Cleans and stores the extracted data into a structured CSV dataset for analysis.
- Designed for speed, simplicity, and reusability, ensuring consistent results across runs.

A simple flow diagram:

```
Website → HTML Fetch (Requests) → Parse Data (BeautifulSoup) → Clean & Store (CSV)
```

Best practice note: In production environments, respect robots.txt directives and platform terms of service. This private testbed operates independently with no external dependencies.

1. Applications of Web Scraping

- Price monitoring and comparison (e-commerce)
- Sentiment analysis from social media
- Job listing aggregation
- Data collection for machine learning
- Academic or market research data gathering

2. Challenges & Limitations

- Dynamic websites (JavaScript-rendered content)
- Captchas and anti-bot mechanisms
- Rate limiting and IP blocking
- Legal and ethical boundaries (robots.txt, GDPR)

3. Future Improvements

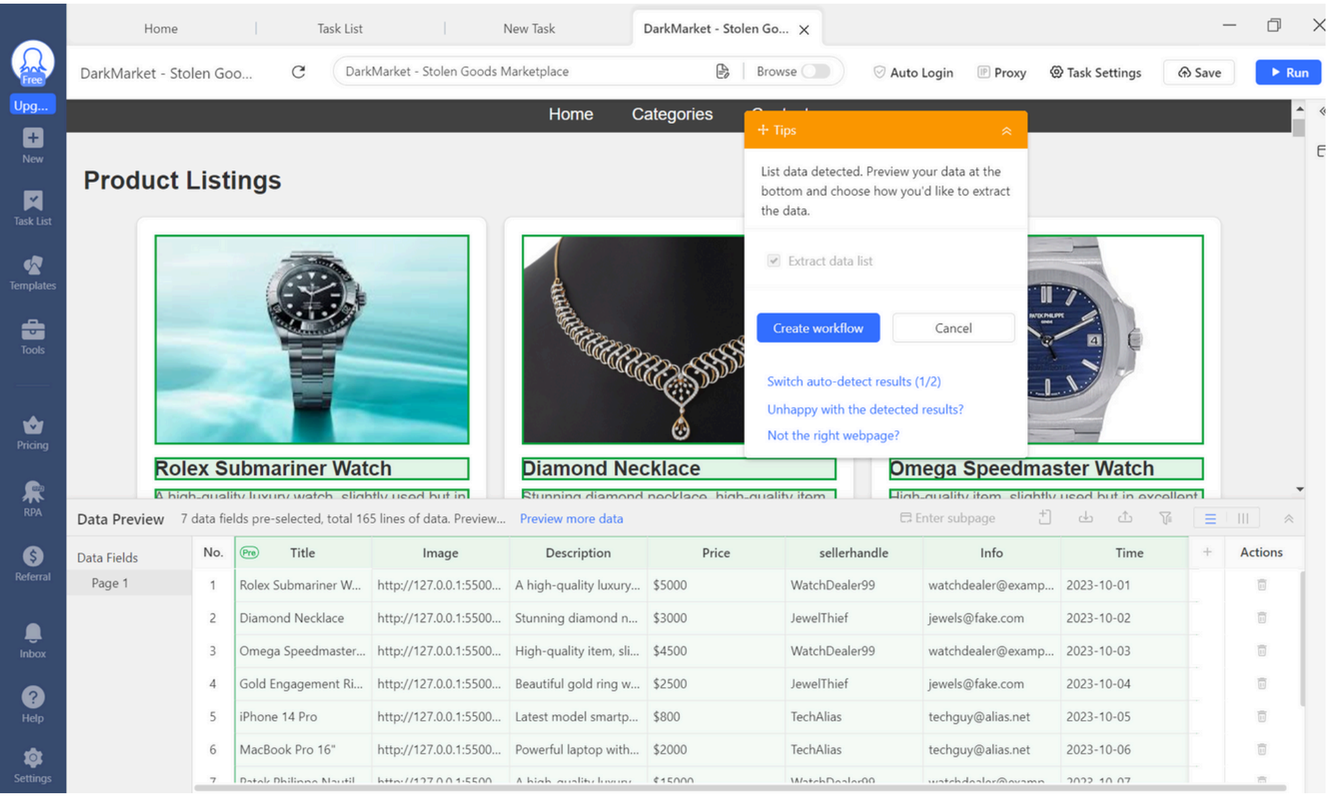
- Automating with a scheduler (cron job)
- Integrating NLP for analyzing scraped text
- Using cloud-based scraping (AWS Lambda, Google Cloud Functions)

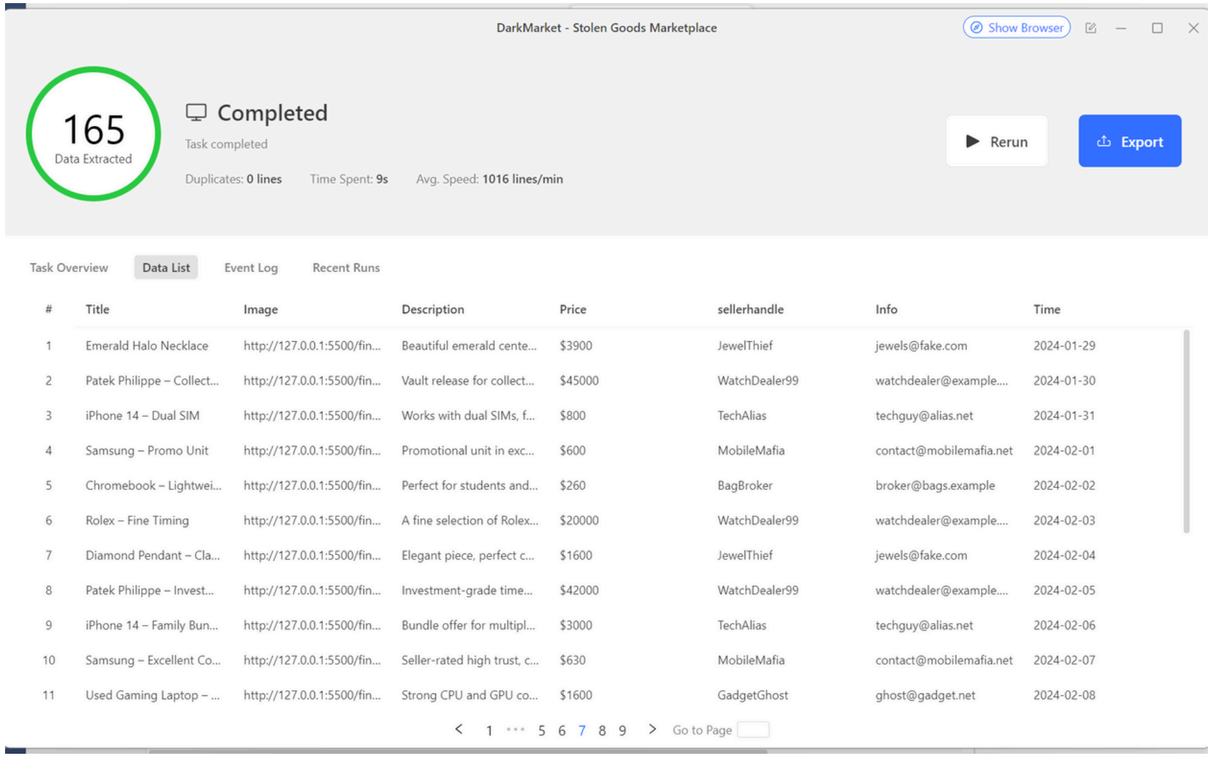
4. Advanced Tools

- Selenium / Playwright – handles JavaScript-heavy sites
- API scraping – more reliable than HTML parsing if available
- octoparse- no code web scraping tool

octoparse tool:

You can open any website inside its built-in browser, click on the elements you want (like product names, prices, emails, etc.), and it will automatically create a scraping workflow.
You can then run the scraper locally on your computer or in the cloud to extract the data into formats like CSV, Excel, JSON, or send it directly to a database or API.





Red-Flag Indicators

To detect patterns associated with stolen or fraudulent goods, eight red-flag indicators were engineered from the scraped marketplace data.

These features captured both behavioral and content-based irregularities:

- Reused Image – Identifies listings that use the same product image across multiple sellers, indicating duplicate or fake listings.
- Shared Contact – Detects repeated contact details (emails or phone numbers) across different sellers, hinting at coordinated or false identities.
- Suspicious Phrase – Flags risky text such as “urgent sale,” “no bill,” “pickup only” that may imply hidden ownership or unverifiable goods.
- Price Outlier – Captures abnormally low-priced items that fall below 50% of the typical product median, signaling possible stolen goods.
- Seller Burst (1-hour / 6-hour) – Highlights sellers posting several listings within short time spans, a sign of automation or bulk dumping.
- High-Velocity Seller – Identifies sellers posting at unusually high daily frequencies, suggesting organized reselling activity.
- Near-Duplicate Description – Detects copied or lightly modified listing descriptions that often appear in fraudulent or spammy posts.

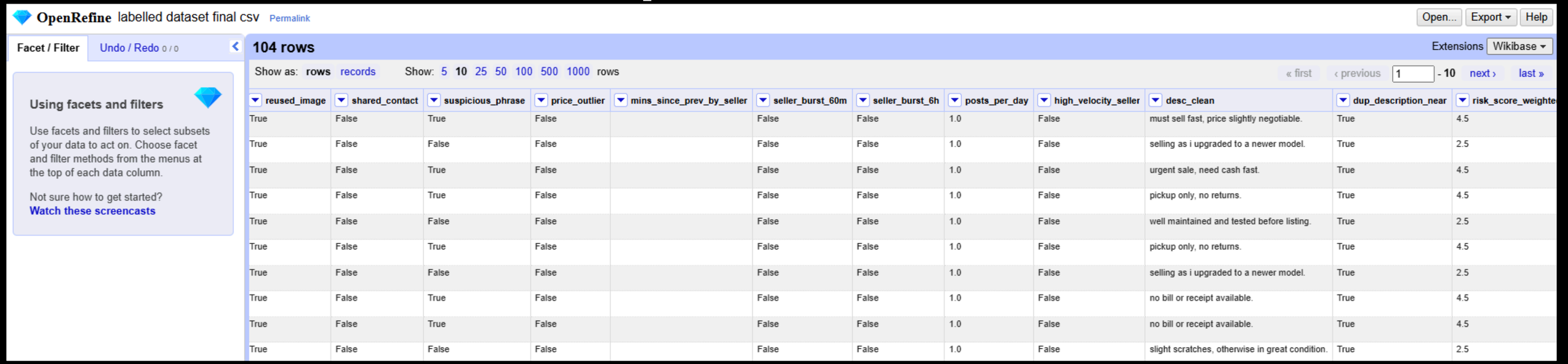
OpenRefine tool

- **OpenRefine** is a free, open-source tool for cleaning and organizing messy data.
- It works like a spreadsheet but is more powerful for data transformation.
- Helps detect and fix inconsistencies, duplicates, and formatting errors.
- Supports data import and export from various formats like CSV, Excel, JSON, and databases.
- Commonly used for data preparation before analysis or machine learning tasks.
- It allows advanced data filtering and clustering to quickly find hidden patterns or errors.
- Can connect to web APIs for fetching or reconciling data with external sources.
- Uses powerful **scripting languages (GREL, Python, or Clojure)** for custom data transformations.

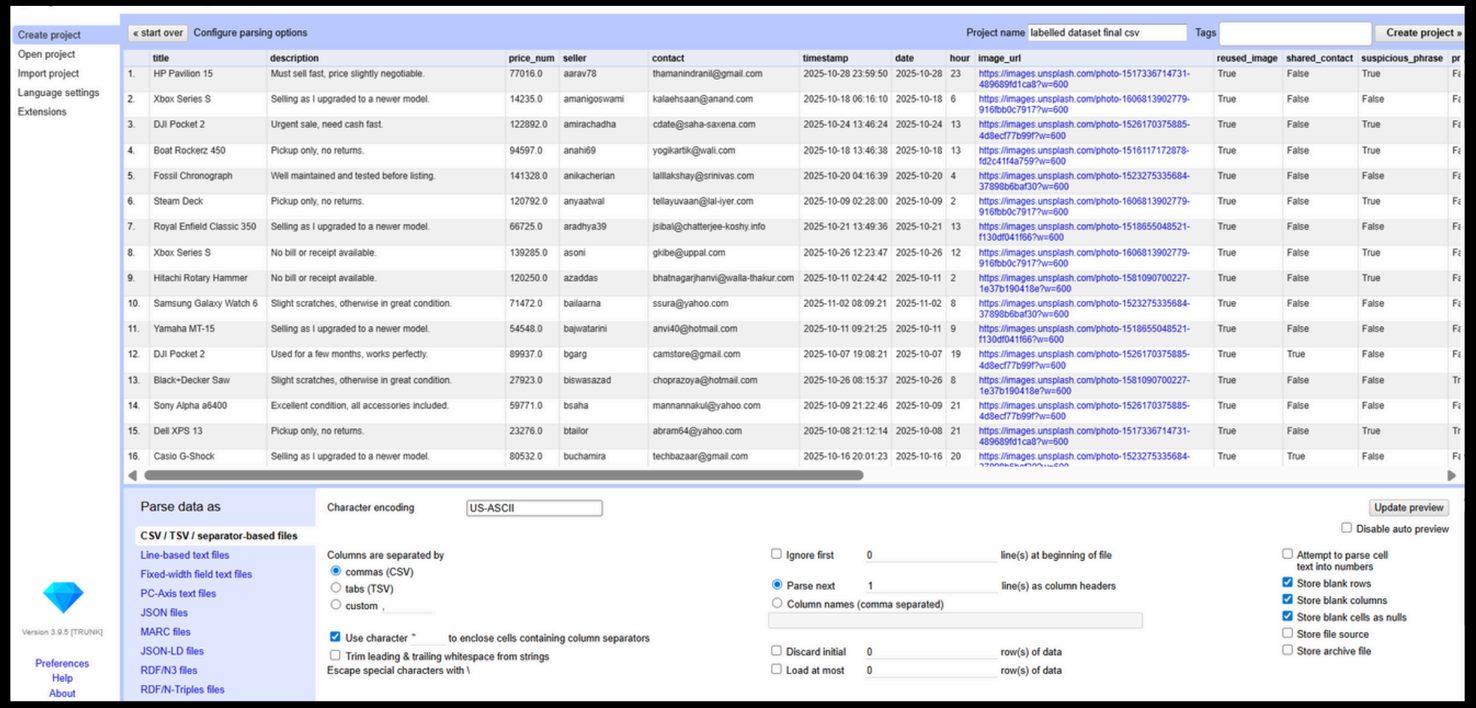
Limitations:-

- OpenRefine cannot handle very large datasets efficiently due to memory limitations.
- It lacks real-time collaboration or multi-user editing features.
- Some advanced operations require knowledge of GREL or scripting, which may be difficult for beginners.

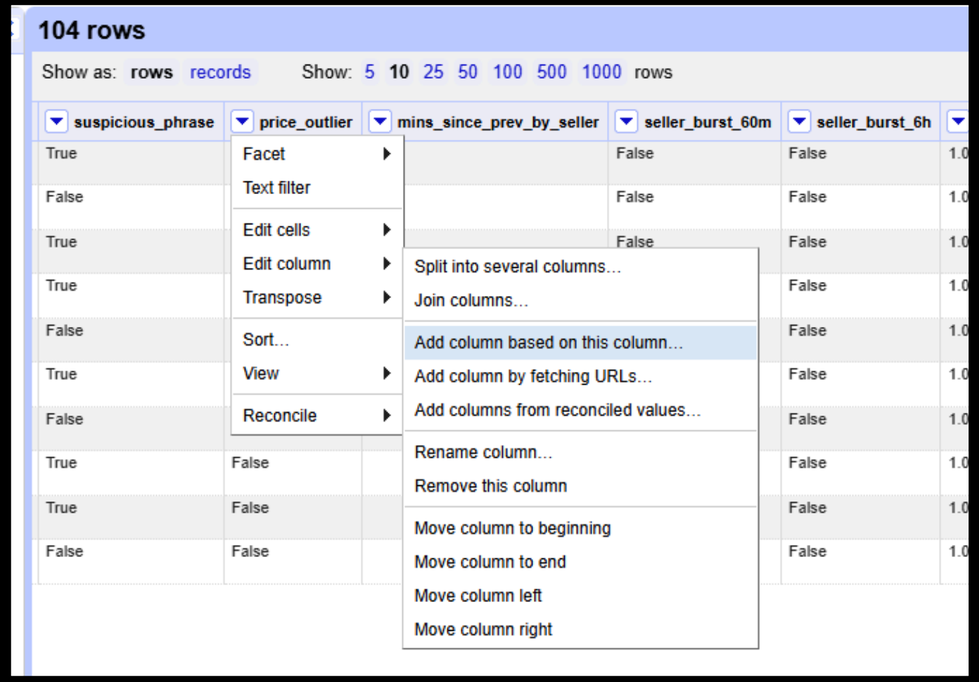
OpenRefine tool



reused_image	shared_contact	suspicious_phrase	price_outlier	mins_since_prev_by_seller	seller_burst_60m	seller_burst_6h	posts_per_day	high_velocity_seller	desc_clean	dup_description_near	risk_score_weighte
True	False	True	False		False	False	1.0	False	must sell fast, price slightly negotiable.	True	4.5
True	False	False	False		False	False	1.0	False	selling as i upgraded to a newer model.	True	2.5
True	False	True	False		False	False	1.0	False	urgent sale, need cash fast.	True	4.5
True	False	True	False		False	False	1.0	False	pickup only, no returns.	True	4.5
True	False	False	False		False	False	1.0	False	well maintained and tested before listing.	True	2.5
True	False	True	False		False	False	1.0	False	pickup only, no returns.	True	4.5
True	False	False	False		False	False	1.0	False	selling as i upgraded to a newer model.	True	2.5
True	False	True	False		False	False	1.0	False	no bill or receipt available.	True	4.5
True	False	True	False		False	False	1.0	False	no bill or receipt available.	True	4.5
True	False	False	False		False	False	1.0	False	slight scratches, otherwise in great condition.	True	2.5

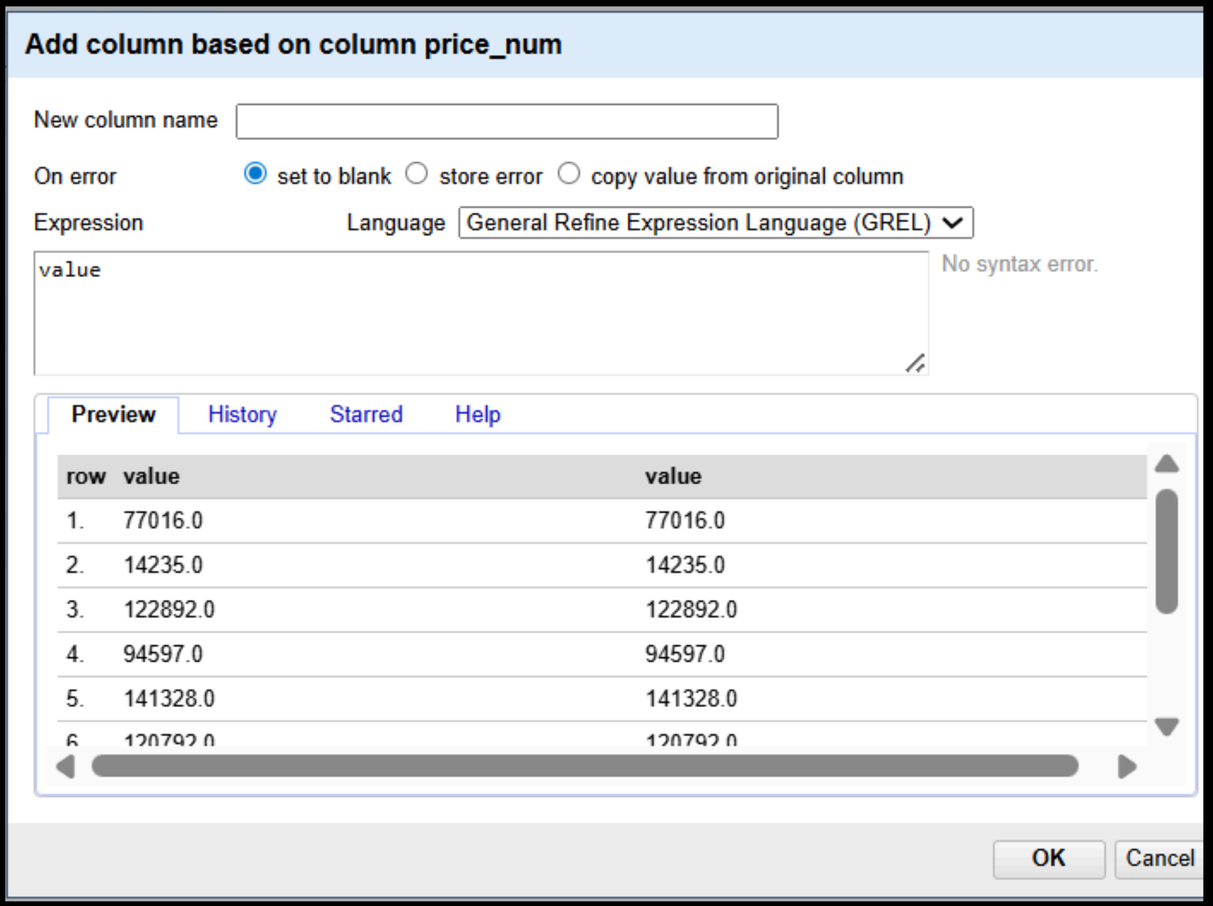


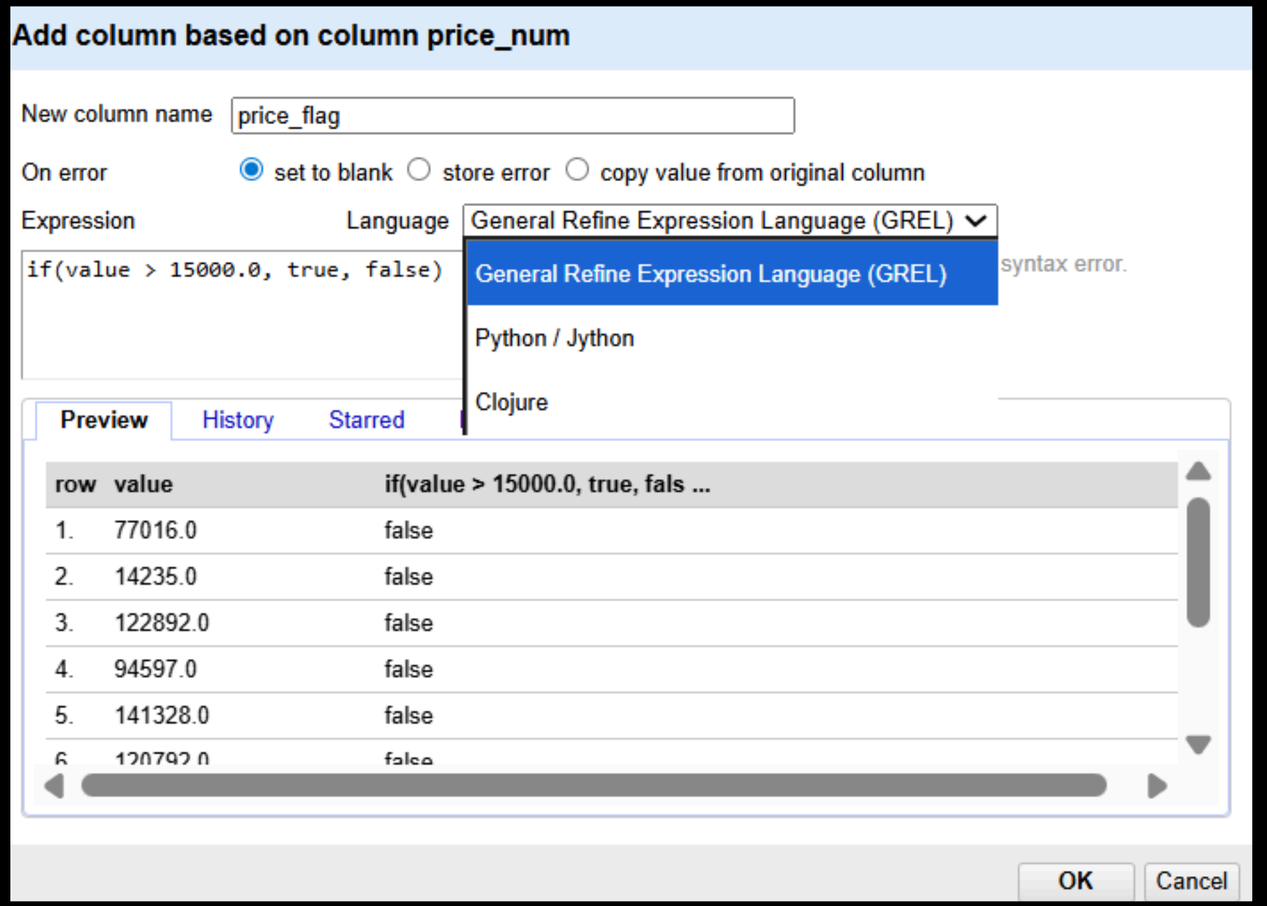
	title	description	price_num	seller	contact	timestamp	date	hour	image_url	reused_image	shared_contact	suspicious_phrase	pr
1.	HP Pavilion 15	Must sell fast, price slightly negotiable.	77016.0	aarav78	thamanindranil@gmail.com	2025-10-28 23:59:50	2025-10-28	23	https://images.unsplash.com/photo-1517336714731-489689fd1ca8?w=600	True	False	True	Fa
2.	Xbox Series S	Selling as I upgraded to a newer model.	14235.0	amanigoswami	kalaehsaan@anand.com	2025-10-18 06:16:10	2025-10-18	6	https://images.unsplash.com/photo-1606813902779-916fbb0c7917?w=600	True	False	False	Fa
3.	DJI Pocket 2	Urgent sale, need cash fast.	122892.0	amirachadha	cdate@saha-saxena.com	2025-10-24 13:46:24	2025-10-24	13	https://images.unsplash.com/photo-1526170375885-4d8ecf77b99f?w=600	True	False	True	Fa
4.	Boat Rockerz 450	Pickup only, no returns.	94597.0	anahi69	yogikartik@wali.com	2025-10-18 13:46:38	2025-10-18	13	https://images.unsplash.com/photo-1516117172878-fd2c41f4a759?w=600	True	False	True	Fa
5.	Fossil Chronograph	Well maintained and tested before listing.	141328.0	anikacherian	lalllakshay@srinivas.com	2025-10-20 04:16:39	2025-10-20	4	https://images.unsplash.com/photo-1523275335684-37898b6baf30?w=600	True	False	False	Fa
6.	Steam Deck	Pickup only, no returns.	120792.0	anyaatwal	tellayuvaan@lal-iyer.com	2025-10-09 02:28:00	2025-10-09	2	https://images.unsplash.com/photo-1606813902779-916fbb0c7917?w=600	True	False	True	Fa
7.	Royal Enfield Classic 350	Selling as I upgraded to a newer model.	66725.0	aradhya39	jsibal@chatterjee-koshy.info	2025-10-21 13:49:36	2025-10-21	13	https://images.unsplash.com/photo-1518655048521-f130df041f66?w=600	True	False	False	Fa
8.	Xbox Series S	No bill or receipt available.	139285.0	asoni	gkibe@uppal.com	2025-10-26 12:23:47	2025-10-26	12	https://images.unsplash.com/photo-1606813902779-916fbb0c7917?w=600	True	False	True	Fa
9.	Hitachi Rotary Hammer	No bill or receipt available.	120250.0	azaddas	bhatnagarjhanvi@walla-thakur.com	2025-10-11 02:24:42	2025-10-11	2	https://images.unsplash.com/photo-1581090700227-1e37b190418e?w=600	True	False	True	Fa
10.	Samsung Galaxy Watch 6	Slight scratches, otherwise in great condition.	71472.0	bailaarna	ssura@yahoo.com	2025-11-02 08:09:21	2025-11-02	8	https://images.unsplash.com/photo-1523275335684-37898b6baf30?w=600	True	False	False	Fa
11.	Yamaha MT-15	Selling as I upgraded to a newer model.	54548.0	bajwatarini	anvi40@hotmail.com	2025-10-11 09:21:25	2025-10-11	9	https://images.unsplash.com/photo-1518655048521-f130df041f66?w=600	True	False	False	Fa
12.	DJI Pocket 2	Used for a few months, works perfectly.	89937.0	bgarg	camstore@gmail.com	2025-10-07 19:08:21	2025-10-07	19	https://images.unsplash.com/photo-1526170375885-4d8ecf77b99f?w=600	True	True	False	Fa
13.	Black+Decker Saw	Slight scratches, otherwise in great condition.	27923.0	biswasazad	choprazoya@hotmail.com	2025-10-26 08:15:37	2025-10-26	8	https://images.unsplash.com/photo-1581090700227-1e37b190418e?w=600	True	False	False	Tr
14.	Sony Alpha a6400	Excellent condition, all accessories included.	59771.0	bsaha	mannannakul@yahoo.com	2025-10-09 21:22:46	2025-10-09	21	https://images.unsplash.com/photo-1526170375885-4d8ecf77b99f?w=600	True	False	False	Fa
15.	Dell XPS 13	Pickup only, no returns.	23276.0	btailor	abram64@yahoo.com	2025-10-08 21:12:14	2025-10-08	21	https://images.unsplash.com/photo-1517336714731-489689fd1ca8?w=600	True	False	True	Tr
16.	Casio G-Shock	Selling as I upgraded to a newer model.	80532.0	buchamira	techbazaar@gmail.com	2025-10-16 20:01:23	2025-10-16	20	https://images.unsplash.com/photo-1523275335684-	True	True	False	Fa



suspicious_phrase	price_outlier	mins_since_prev_by_seller	seller_burst_60m	seller_burst_6h
True			False	False
False			False	False
True			False	False
True				False
False				False
True				False
False				False
True	False			False
True	False			False
False	False			False

OpenRefine tool





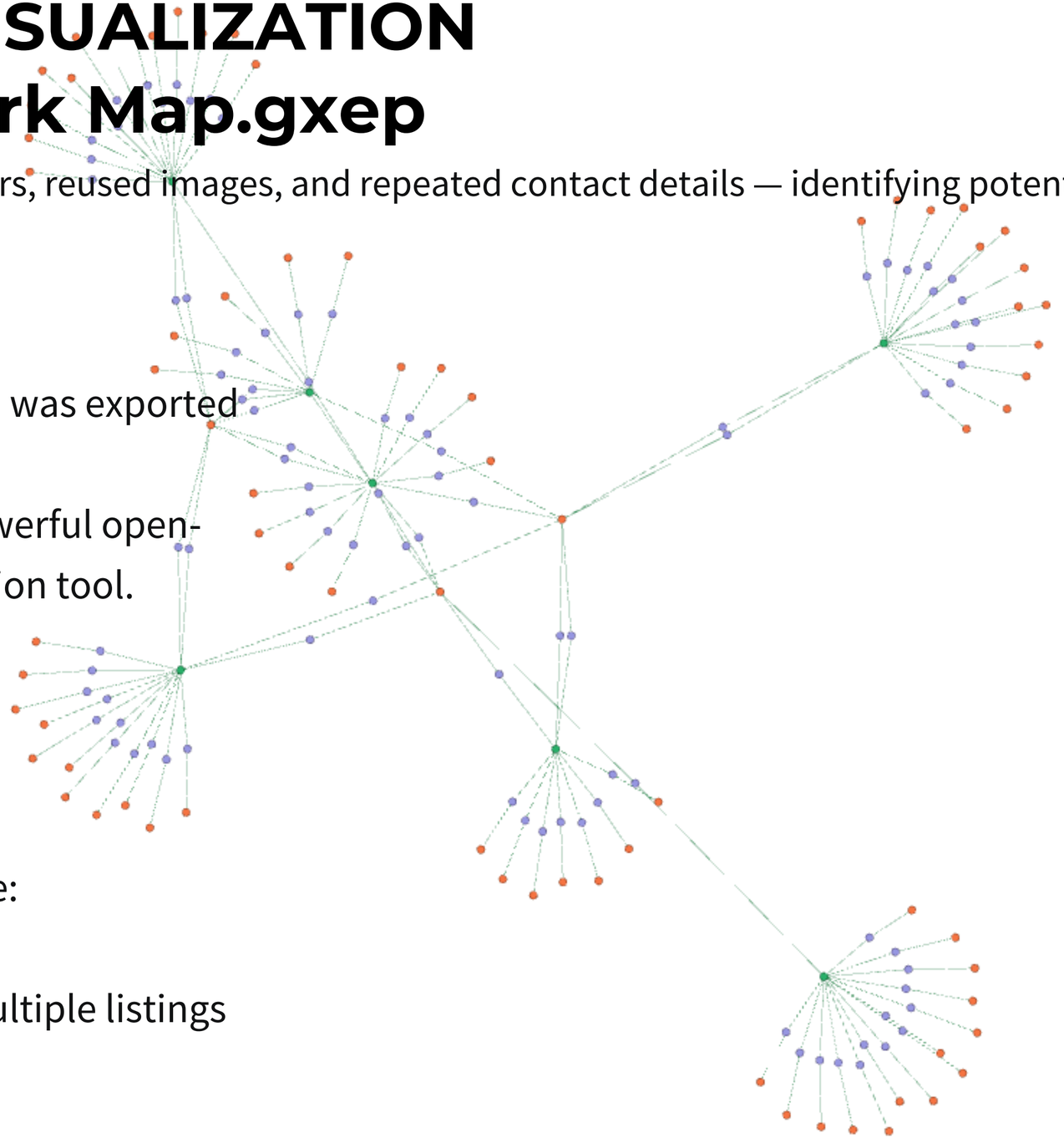
Weighted Risk Score Calculation

To quantify the overall suspiciousness of each listing, a weighted risk-scoring system was designed.
This converts the eight red-flag indicators into a single interpretable metric that measures risk severity.

- Purpose: Combine multiple behavioral and textual warning signs into one numeric score to rank listings by their likelihood of being fraudulent.
- Weighted Approach:
  - The first four indicators (Reused Image, Shared Contact, Suspicious Phrase, Price Outlier) were given double weight (2×) because they directly suggest fraudulent intent.
  - The behavioral indicators (Seller Burst 1h/6h, High-Velocity Seller) were given medium weight (1×) to represent suspicious posting activity.
  - The Near-Duplicate Description indicator was given light weight (0.5×) since repetition alone may not imply fraud.
- Calculation Formula:
- Each listing's Risk Score = Σ (weight × indicator value).
- Example →
- Risk Score = 2*(reused_image + shared_contact + suspicious_phrase + price_outlier) + 1*(seller_burst_60m + seller_burst_6h + high_velocity_seller) + 0.5*(dup_description_near)
- Threshold Selection:
- Listings with a weighted score ≥ 3 were labeled Suspicious (1), while others were considered Normal (0).
- This threshold balanced detection sensitivity and false positives.
- Risk Level Classification:
  - Low Risk: Score < 2
  - Medium Risk: 2 ≤ Score < 3
  - High Risk: Score ≥ 3

GEPHI TOOL FOR VISUALIZATION
Building the Network Map.gxep

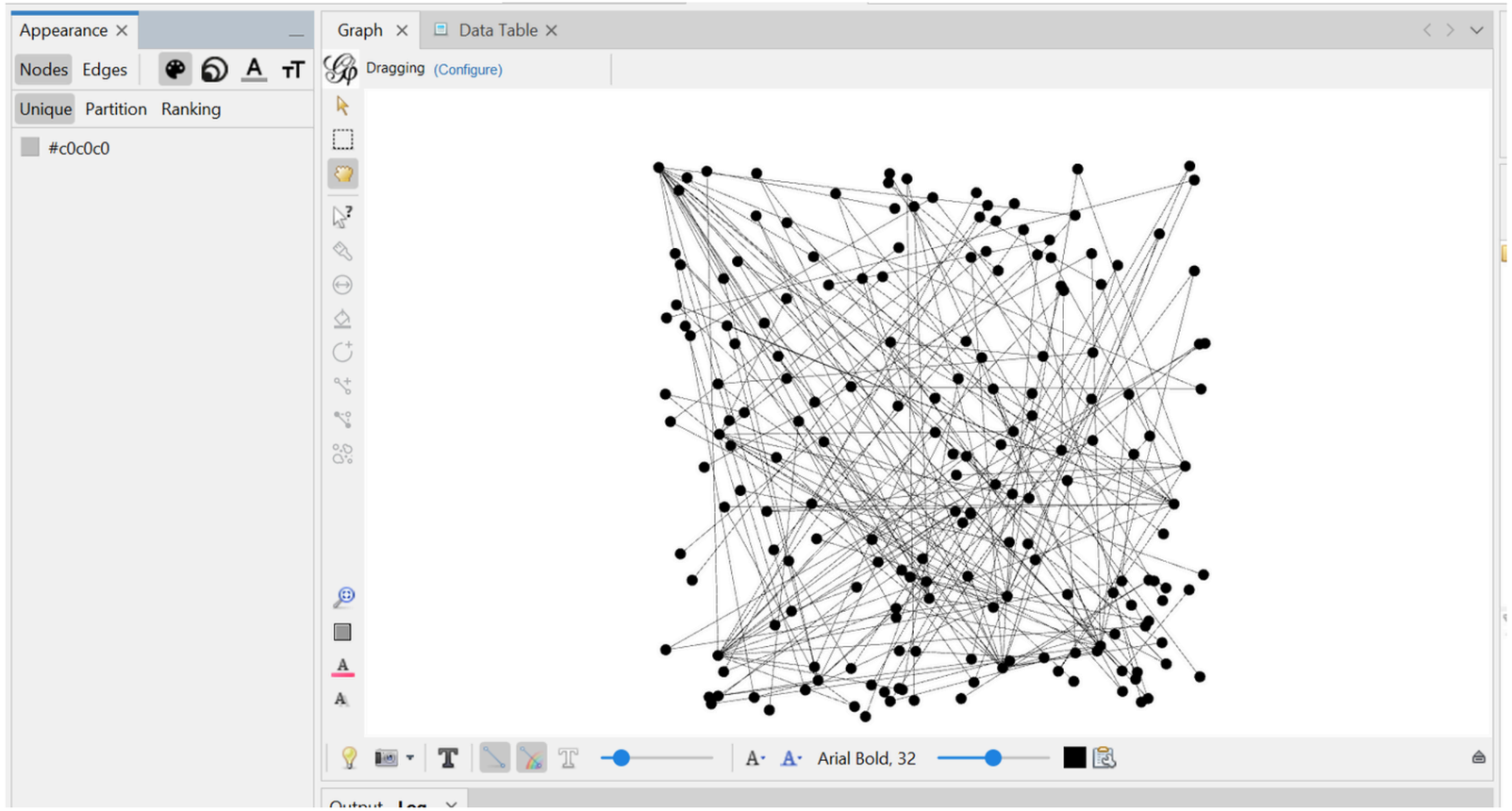
To uncover hidden connections between sellers, reused images, and repeated contact details — identifying potential illicit activity clusters on the marketplace.



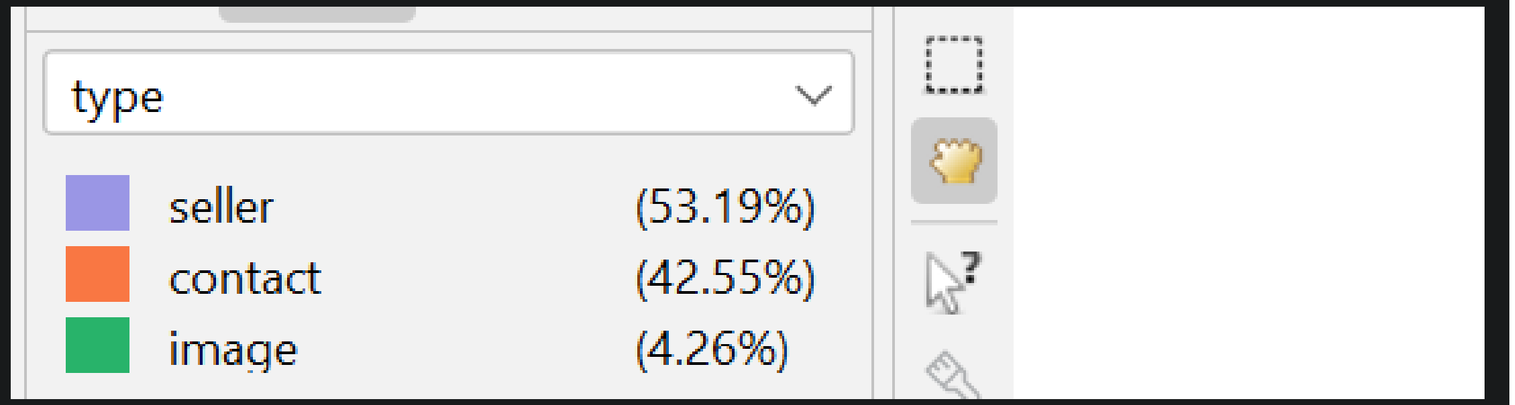
Methodology

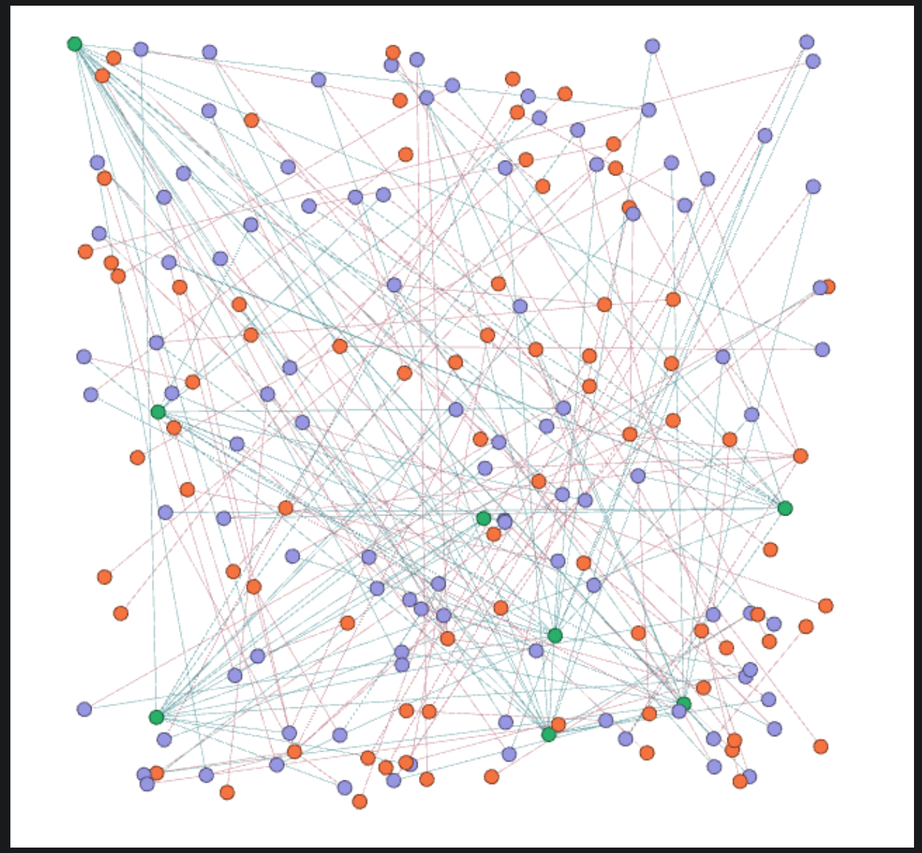
- The scraped data (from BeautifulSoup) was exported as a .gexf file.
- The file was imported into Gephi, a powerful open-source network analysis and visualization tool.
- Nodes represented:
  - 🧍 Sellers
  - ☎️ Contact Numbers
  - 🖼️ Image URLs
- Edges (connections) were drawn where:
  - The same image was reused
  - The same contact was linked to multiple listings

Network Graph without Filters or Legend

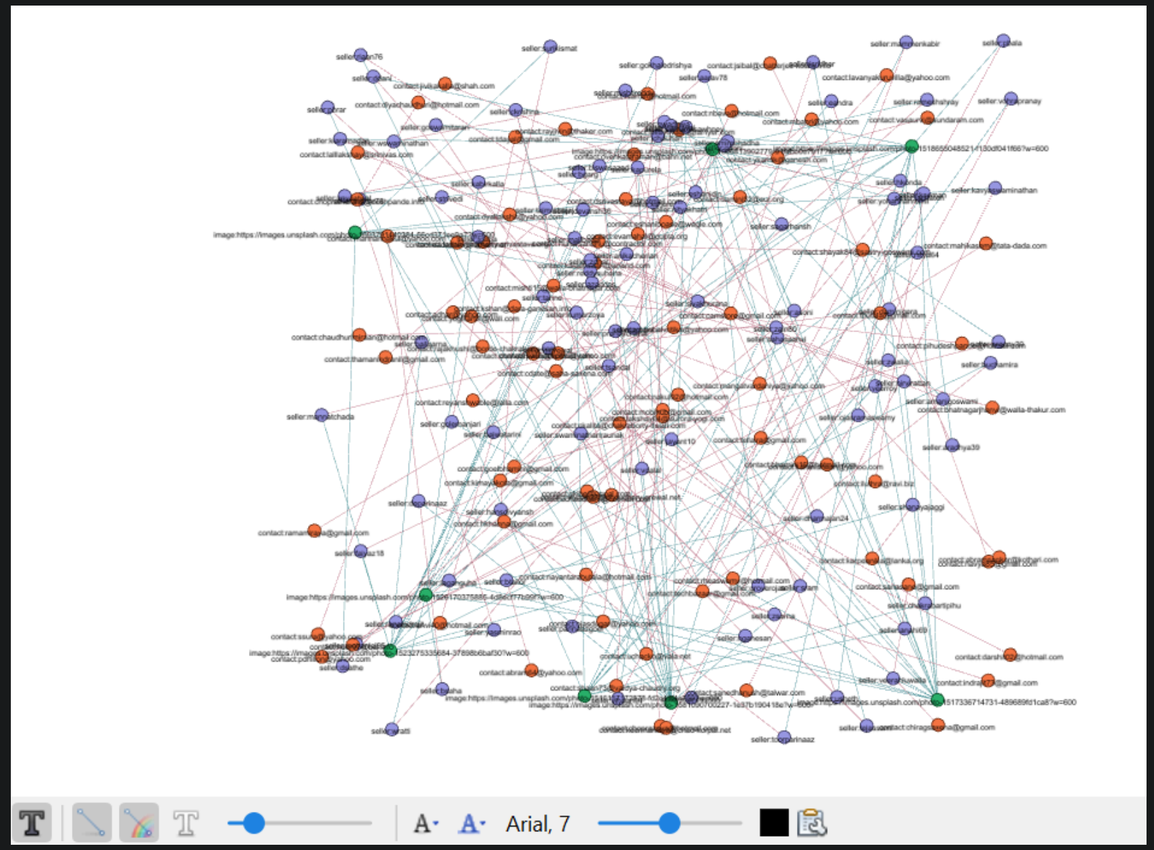


Visualizing Relationships by Type: Sellers, Contacts, and Images

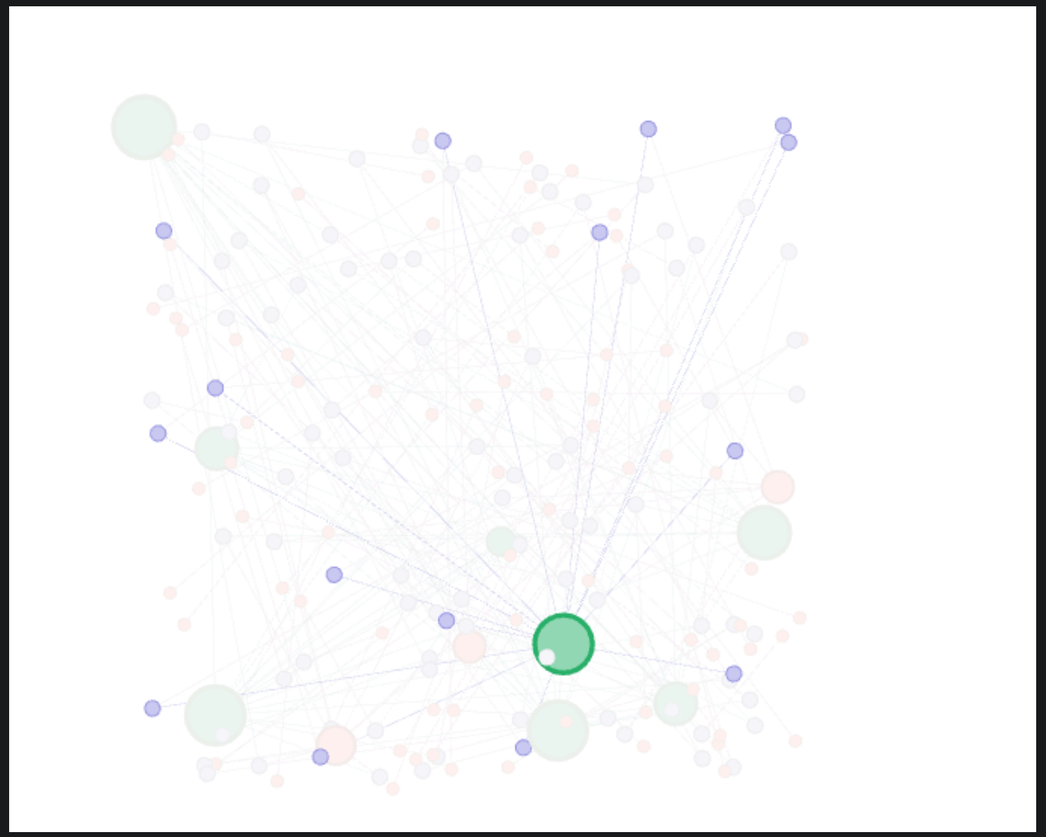




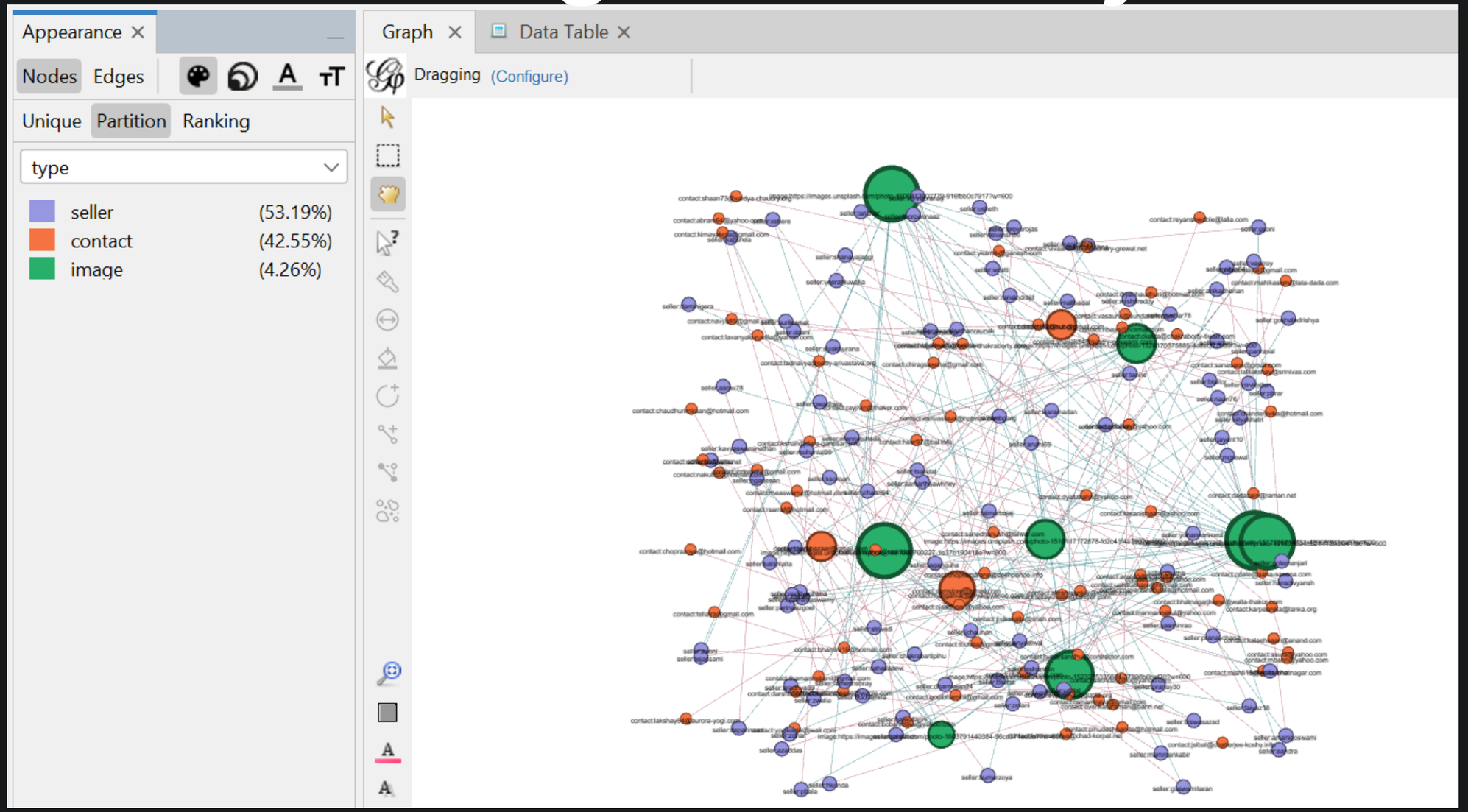
ADDING LABELS TO NODES.



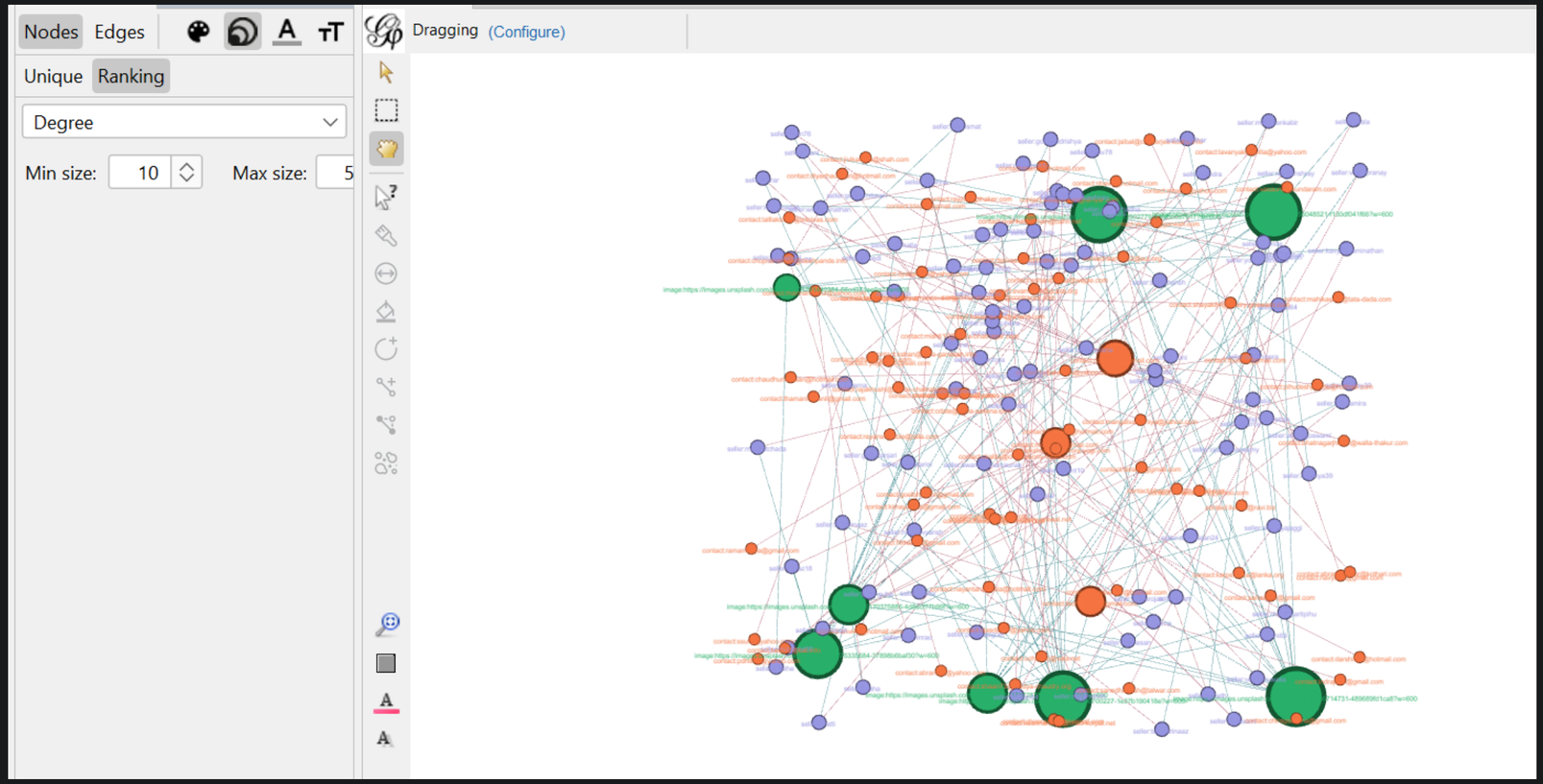
Hover on a node to get its connections.



seller_network.gexf (SIZE=degree centrality)

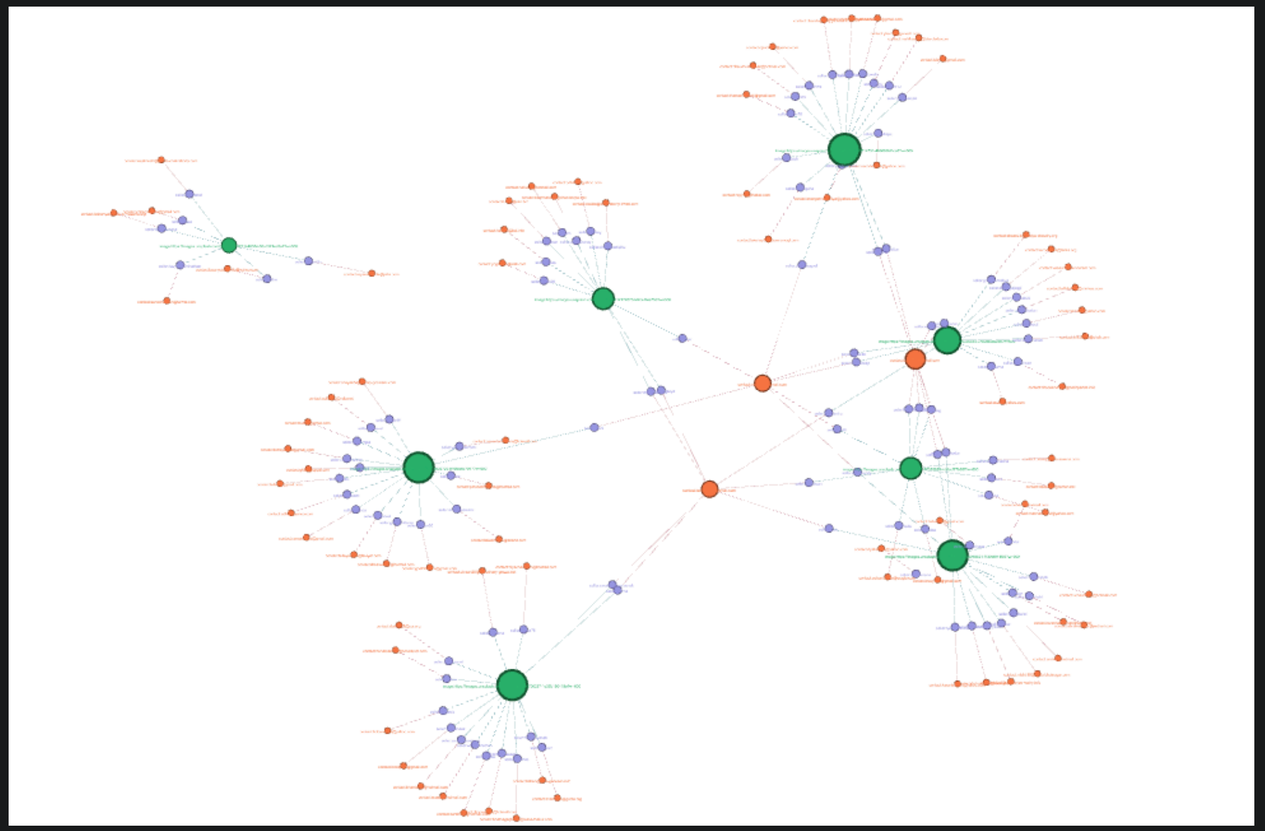


LABEL COLOR IS MADE SAME AS NODE COLOR (BY CLICKING ON “A” IN TOP LEFT AND APPLY)

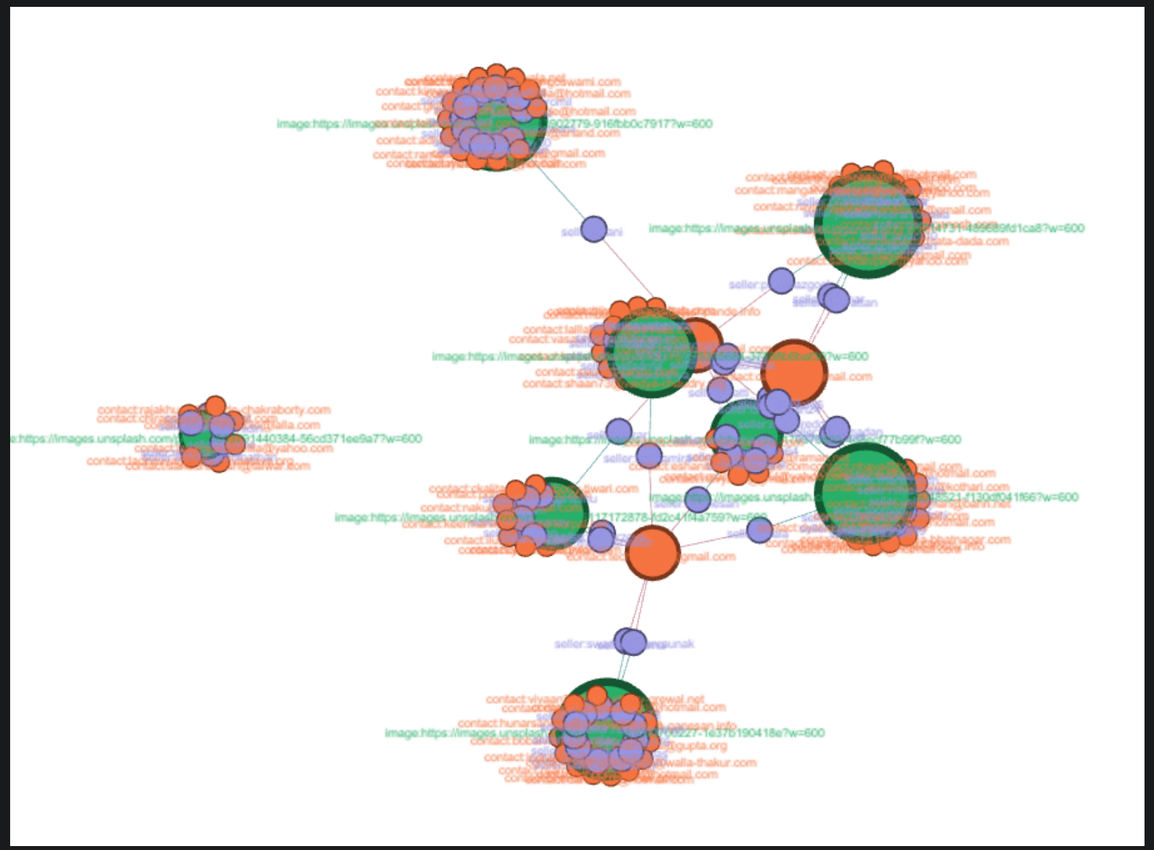


EXPLORING DIFFERENT INBUILD LAYOUTS IN THE TOOL.

YIFAN HU LAYOUT:



FORCE ATLAS LAYOUT:



Hybrid Clustering Approach for Suspicious Seller Detection

The objective of this module is to identify patterns of coordinated or fraudulent behaviour among online marketplace sellers.
Instead of analysing each listing independently, the approach looks for relationships between listings — either through textual similarity or shared digital identifiers — to uncover hidden connections that may indicate duplicate or fake seller accounts

🔴 Text-Based Clustering

focuses on the linguistic and semantic similarity of product descriptions.
Listings that describe products in nearly identical language are grouped together, even when the wording is slightly altered.
This helps expose reused or templated descriptions across different accounts.

🔴 Graph-Based Clustering

models the relationships between sellers as a social-network graph.
Connections are created whenever two sellers share the same contact information or reuse the same product image, signalling possible coordination or duplicate identities.

By combining textual and relational evidence, the system detects potential multi-account misuse, fake listings, and unauthorised reselling. It provides a structured, data-driven way to visualise and verify suspicious clusters that would be difficult to notice manually.

Text-Based Clustering (Semantic Similarity)

Text-based clustering aims to detect listings that use similar or identical product descriptions across different sellers.
Even when words or sentence structures vary, sellers may reuse the same underlying text patterns — often a strong indicator of coordinated copied behaviour.

Methodology

- Product descriptions are first transformed into semantic representations using Sentence-BERT, a deep learning model trained to understand sentence meaning rather than just keywords.
- The similarity between every pair of listings is then measured using cosine similarity, which quantifies how closely their meanings align.
- Listings with similarity scores above a predefined threshold (e.g., 0.82) are grouped together into clusters, representing semantically identical or near-identical descriptions.
- Each cluster represents a potential case of cross-seller text reuse or copy-paste listings.

Interpretation

- When multiple sellers use nearly identical descriptions for similar products, it may indicate:
  - Template-based fraudulent posting, where the same seller operates multiple accounts.
  - Content plagiarism, where legitimate sellers copy text from others to appear authentic.
  - Automated listing generation, where bots replicate product descriptions.

Outcome:

- **The approach identified multiple semantic text clusters containing listings that are contextually similar despite minor rewording.**
- **These clusters highlight reused content patterns that can serve as an early warning for fraudulent or duplicate listings.**

Text-Based Clustering-Output

Found 17 suspicious text clusters (cross-seller reuse)

	Cluster ID	Seller	Title	Description
0	0	aarav78	HP Pavilion 15	Must sell fast, price slightly negotiable.
1	0	amirachadha	DJI Pocket 2	Urgent sale, need cash fast.
2	0	eandra	Sony Alpha a6400	Urgent sale, need cash fast.
3	1	amanigoswami	Xbox Series S	Selling as I upgraded to a newer model.
4	1	aradhya39	Royal Enfield Classic 350	Selling as I upgraded to a newer model.
5	1	bajwatarini	Yamaha MT-15	Selling as I upgraded to a newer model.
6	2	anahi69	Boat Rockerz 450	Pickup only, no returns.
7	2	anyaatwal	Steam Deck	Pickup only, no returns.
8	2	btailor	Dell XPS 13	Pickup only, no returns.
9	3	anikacherian	Fossil Chronograph	Well maintained and tested before listing.
10	3	ckrishna	Black+Decker Saw	Well maintained and tested before listing.
11	3	golemanjari	HP Pavilion 15	Well maintained and tested before listing.
12	6	bgarg	DJI Pocket 2	Used for a few months, works perfectly.
13	6	kabirkalla	Canon EOS 250D	Used for a few months, works perfectly.
14	6	maninishith	Honda Shine	Used for a few months, works perfectly.

Observations

The text-based clustering grouped listings with semantically similar descriptions, revealing sellers who use almost identical product texts.

Each cluster represents potential reused or templated listings — for example, multiple sellers using the same sales pitch like “**Selling as I upgraded to a newer model**” or “**Pickup only, no returns**”.

This helps flag coordinated or duplicate seller behavior across the marketplace.

Graph-Based Clustering (Seller Relationship Network)

Graph-based clustering extends the analysis beyond textual data by modelling relationships between sellers based on shared attributes. This method visualises how sellers are connected through shared digital identifiers, enabling the detection of groups that operate as coordinated entities.

Methodology

- Each seller is represented as a node in a network graph.
- Edges (connections) are created whenever two sellers share:
  - The same contact information (e.g., phone number or email), or
  - The same product image reused across listings.
- These relationships form a web of interconnected sellers.
- Using a modularity-based community detection algorithm, the network is divided into communities (clusters) where sellers are densely interconnected.
- Each community thus represents a potential group of accounts linked through shared identifiers.

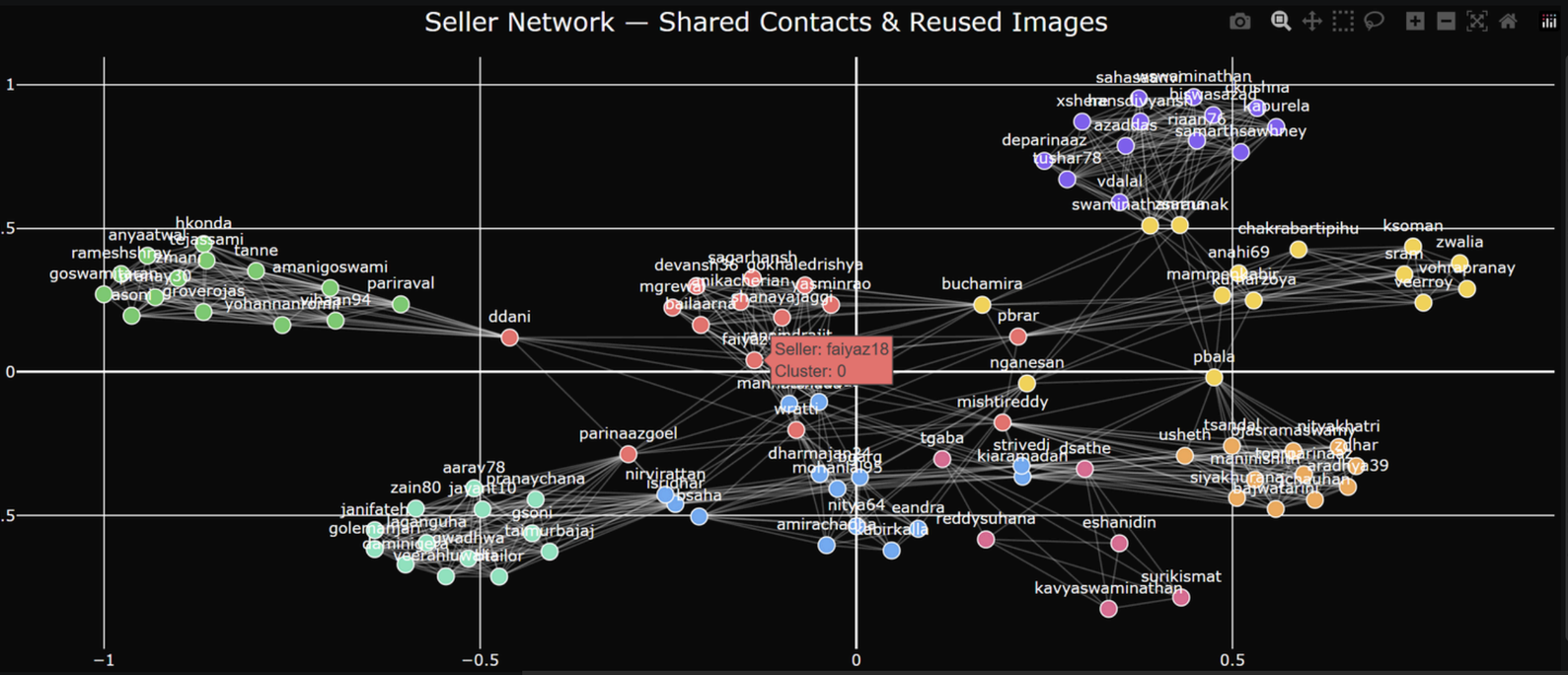
Interpretation

- Sellers within the same cluster are likely managed by the same individual or team, even if their public account names differ.
- Shared contacts suggest operational overlap, while reused images may indicate duplicate listings for the same item across different accounts.
- The resulting network provides a visual map of coordinated behaviour, where tightly grouped nodes highlight possible fraud rings.

Outcome:

- **Several distinct seller communities were detected, each representing possible multi-account networks or duplicate resellers.**
- **These clusters can be explored visually to trace the contact or image links that connect sellers.**

Graph-Based Clustering-Output



Summary

The graph-based clustering revealed several interconnected seller groups linked through shared contacts and reused product images. Each cluster likely represents coordinated accounts managed by the same individual or team.

For example, clusters sharing the same email or image URL indicate possible multi-account activity or duplicate listings

example, Cluster 0 is linked by the contact mobihub@gmail.com, while others reuse identical product images.

This confirms how graph-based clustering helps uncover hidden seller networks operating under multiple accounts.

```
🔹 Cluster 0 has 15 Sellers
Shared Contacts: ['mobihub@gmail.com']
Shared Images: ['https://images.unsplash.com/photo-1523275335684-37898b6baf30?w=600']

🔹 Cluster 1 has 14 Sellers
Shared Contacts: ['camstore@gmail.com']
Shared Images: ['https://images.unsplash.com/photo-1526170375885-4d8ecf77b99f?w=600', 'https://images.unsplash.com/photo-1517336714731-489689fd1ca8?w=600',

🔹 Cluster 2 has 14 Sellers
Shared Contacts: ['ramamiraya@gmail.com']
Shared Images: ['https://images.unsplash.com/photo-1606813902779-916fbb0c7917?w=600']

🔹 Cluster 3 has 14 Sellers
Shared Contacts: ['techbazaar@gmail.com', 'iluthra@ravi.biz']
Shared Images: ['https://images.unsplash.com/photo-1516117172878-fd2c41f4a759?w=600', 'https://images.unsplash.com/photo-1581090700227-1e37b190418e?w=600']

🔹 Cluster 4 has 13 Sellers
Shared Images: ['https://images.unsplash.com/photo-1581090700227-1e37b190418e?w=600']

🔹 Cluster 5 has 13 Sellers
Shared Images: ['https://images.unsplash.com/photo-1517336714731-489689fd1ca8?w=600']

🔹 Cluster 6 has 11 Sellers
Shared Images: ['https://images.unsplash.com/photo-1518655048521-f130df041f66?w=600']

🔹 Cluster 7 has 6 Sellers
Shared Images: ['https://images.unsplash.com/photo-1603791440384-56cd371ee9a7?w=600']
```

Suspicious Listing Detection using Machine Learning

We built a machine-learning–based detection tool to automatically assess the likelihood of a listing being suspicious.
Using the scraped dataset, the model evaluates indicators such as reused images, shared contacts, suspicious phrases, and pricing anomalies to generate a percentage “risk score” for each listing.

Model Overview

Algorithm: Isolation Forest (Unsupervised Anomaly Detection)
Framework: scikit-learn
Input Features:

- Price value (price_num)
- Reused image indicator (reused_image)
- Shared contact indicator (shared_contact)
- Price anomaly flag (price_outlier)
- Seller posting frequency (posts_per_day)
- Seller burst activity (seller_burst_60m, seller_burst_6h)
- High-velocity seller flag (high_velocity_seller)
- Duplicate description similarity (dup_description_near)
- Output: Suspiciousness Score (0–100%) — representing how different a seller’s behavior is from the norm.

Results Interpretation

Risk Level	Range (%)	Meaning
Low	0–40	Likely normal seller
Medium	41–70	Needs manual review
High	71–100	Strongly suspicious — reuse/contact overlap detected

Unsupervised ML for Suspicious Seller Detection

Isolation Forest: Detecting Suspicious Seller Behavior

Key Features Used:

- Price anomalies (price_num, price_outlier)
- Behavioral metrics (seller_burst_60m, posts_per_day, high_velocity_seller)
- Fraud indicators (reused_image, shared_contact, dup_description_near)

Model Overview

- In real marketplaces, ground-truth fraud labels are unavailable.
- Hence, a supervised model (like XGBoost) isn't suitable.
- To overcome this, we used Isolation Forest, an unsupervised anomaly detection algorithm that identifies outliers based on behavioral deviations.
- **Goal: Automatically flag sellers who behave differently from normal patterns.**

Technical Workflow

- Data preprocessing and feature scaling
- Train Isolation Forest (contamination = 0.1)
- Compute anomaly scores → convert to 0–100 suspiciousness percent
- Assign risk levels → Low, Medium, High

Unsupervised ML for Suspicious Seller Detection-Ouput

Isolation Forest: Detecting Suspicious Seller Behavior

Sample of Suspicious Listings:

	seller	price_num	reused_image	shared_contact	suspicious_percent	ml_risk_category
0	aarav78	77016.0	True	False	90.28	High
1	amanigoswami	14235.0	True	False	63.69	Medium
2	amirachadha	122892.0	True	False	83.88	High
3	anahi69	94597.0	True	False	93.19	High
4	anikacherian	141328.0	True	False	82.49	High
5	anyaatwal	120792.0	True	False	84.12	High
6	aradhya39	66725.0	True	False	94.66	High
7	asoni	139285.0	True	False	68.34	Medium
8	azaddas	120250.0	True	False	83.25	High
9	bailaarna	71472.0	True	False	91.10	High

Observations

The machine learning model assigns a **suspiciousness percentage** to each listing based on indicators like reused images, shared contacts, and pricing anomalies.

Most listings in this sample fall under the **High-risk category (80–95%)**, indicating strong evidence of coordinated or fraudulent behavior.

Only a few listings are rated as **Medium risk**, showing moderate suspicious patterns but not enough overlap for full flagging.

This confirms that the model effectively differentiates between potentially fraudulent sellers and relatively safer ones.

Outcome:

Overall, this demonstrates the system’s capability to automatically identify high-risk sellers and prioritize them for manual verification — reducing human effort and improving fraud detection accuracy.

OSINT Reconnaissance

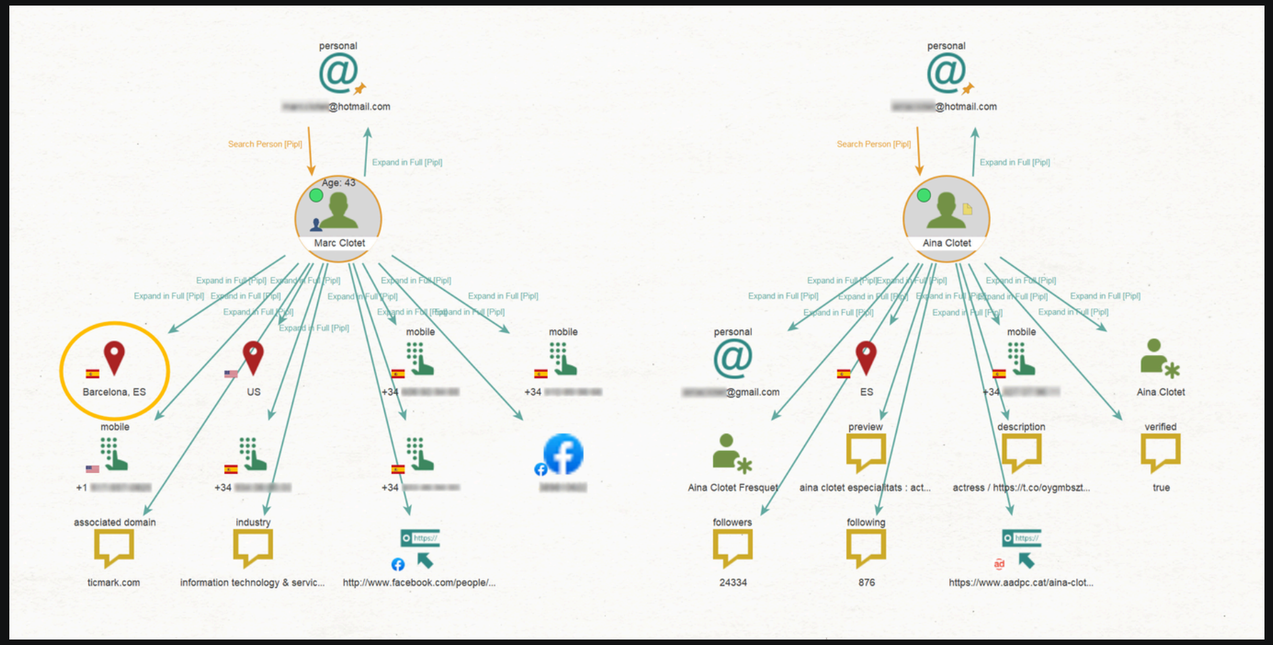
To validate or enrich the suspicious data you scraped (emails, phone numbers, images) using public information sources.

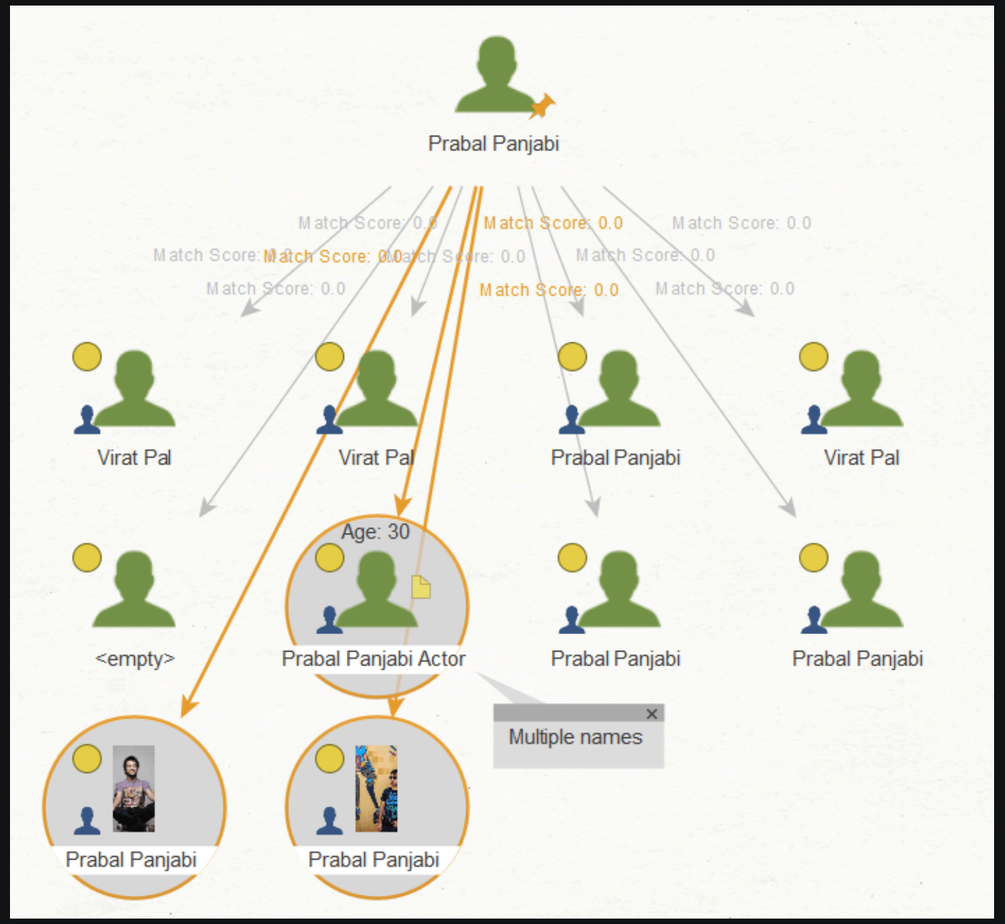
Maltego

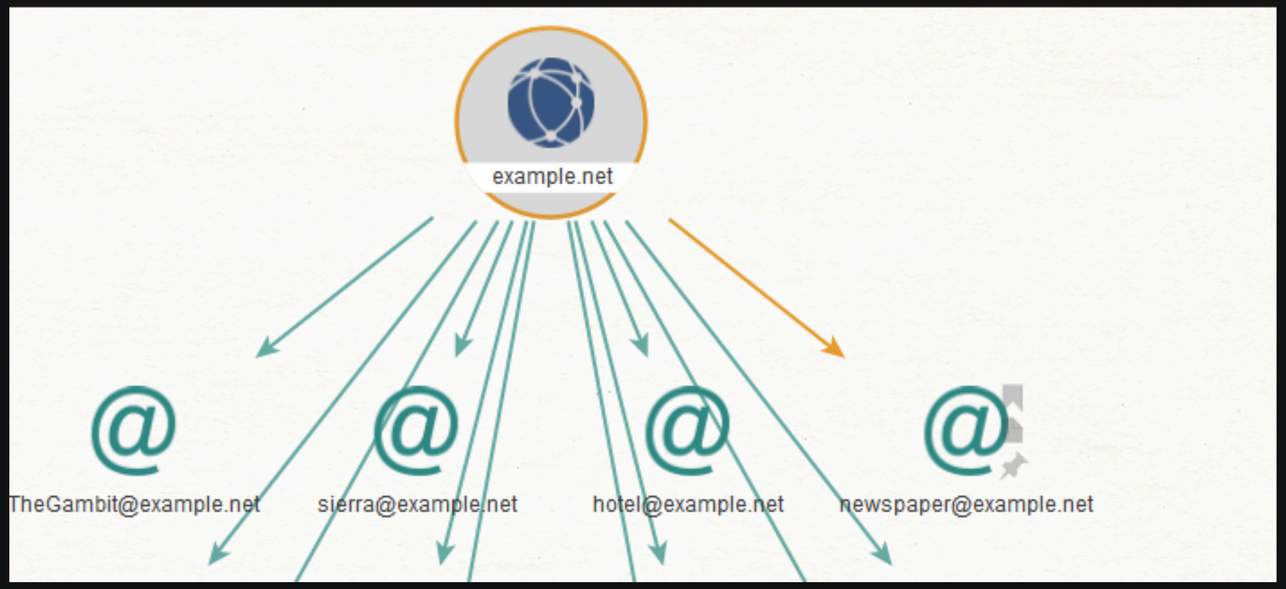
Maltego is an open-source intelligence (OSINT) and link analysis tool used for investigating relationships between digital entities such as emails, domains, and social media accounts. In our project, it helps trace connections between suspicious seller emails and related online footprints, visualizing possible networks of coordinated activity.

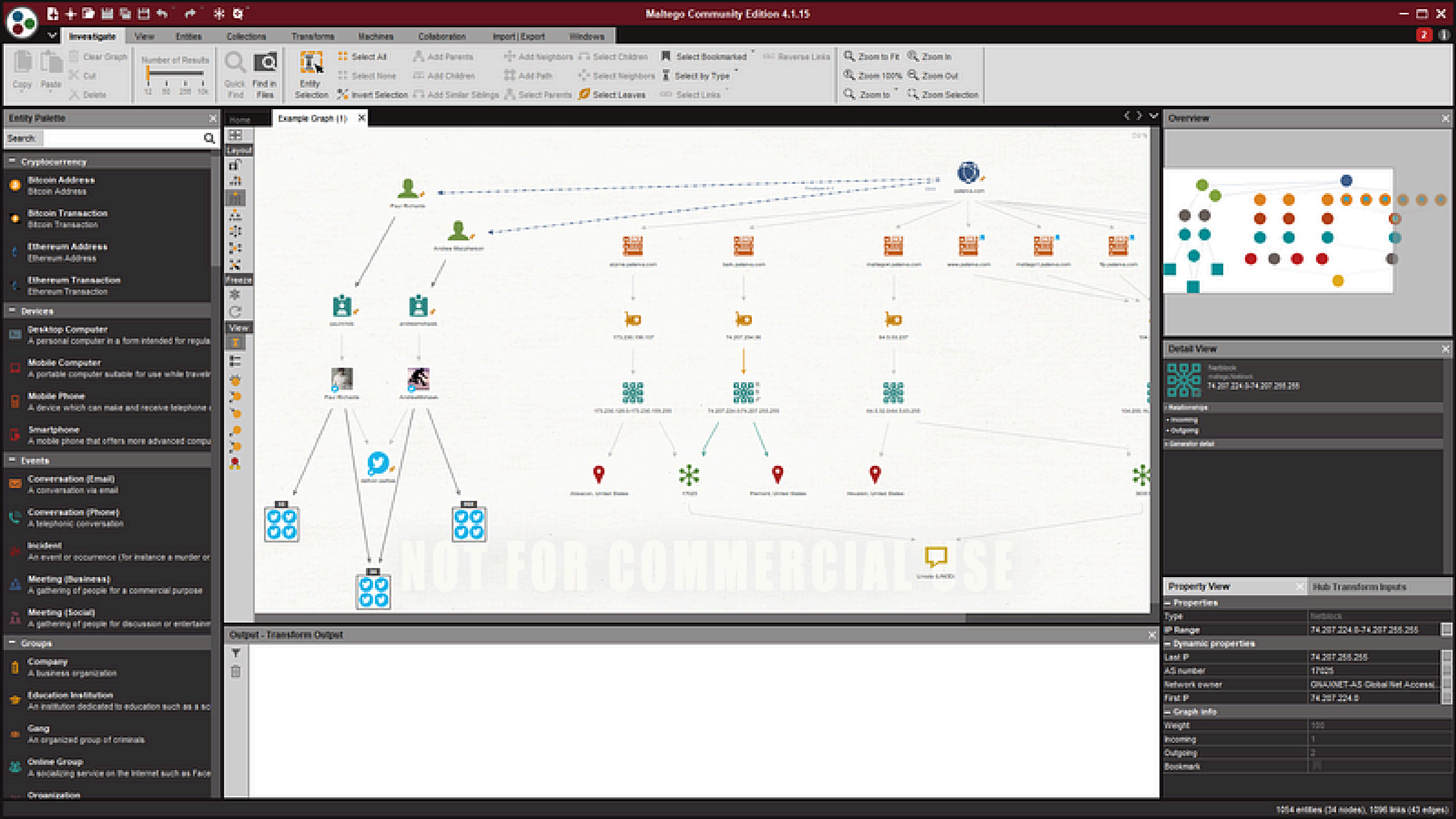
NumVerify

Numverify is a phone number validation and lookup tool that provides details such as carrier, country, and line type. We used it to verify seller contact numbers and identify patterns like reused VoIP or international numbers, which can indicate fake or fraudulent seller accounts.


personal
@hotmail.com
Search Person [Pipl]
Expand in Full [Pipl]
Age: 43
Marc Clotet
Barcelona, ES
US
mobile
+34
mobile
+34
mobile
+1
+34
+34
associated domain
ticmark.com
industry
information technology & servic...
http://www.facebook.com/people/...
personal
@hotmail.com
Aina Clotet
personal
@gmail.com
ES
mobile
+34
Aina Clotet
Aina Clotet Fresquet
preview
aina clotet especialitats : act...
description
actress / https://t.co/oygmbszt...
verified
true
followers
24334
following
876
https://www.aadpc.cat/aina-clot...


Prabal Panjabi
Match Score: 0.0
Virat Pal
Virat Pal
Prabal Panjabi
Virat Pal
<empty>
Age: 30
Prabal Panjabi Actor
Prabal Panjabi
Prabal Panjabi
Multiple names
Prabal Panjabi
Prabal Panjabi


example.net
TheGambit@example.net
sierra@example.net
hotel@example.net
newspaper@example.net


Maltego Community Edition 4.1.15
Investigate
View
Entities
Collections
Transform
Machines
Collaboration
Import | Export
Windows
Entity Palette
Search
Cryptocurrency
Bitcoin Address
Bitcoin Transaction
Ethereum Address
Ethereum Transaction
Devices
Desktop Computer
Mobile Computer
Mobile Phone
Smartphone
Events
Conversation (Email)
Conversation (Phone)
Incident
Meeting (Business)
Meeting (Social)
Groups
Company
Education Institution
Gang
Online Group
Organization
Example Graph (1)
Layout
Freeze
View
Overview
Detail View
Property View
Hub Transform Inputs
Properties
Type
IP Range
Dynamic properties
Last IP
AS number
Network owner
First IP
Graph info
Weight
Incoming
Outgoing
Bookmark
Output - Transform Output
NOT FOR COMMERCIAL USE

Maltego

steps Performed:

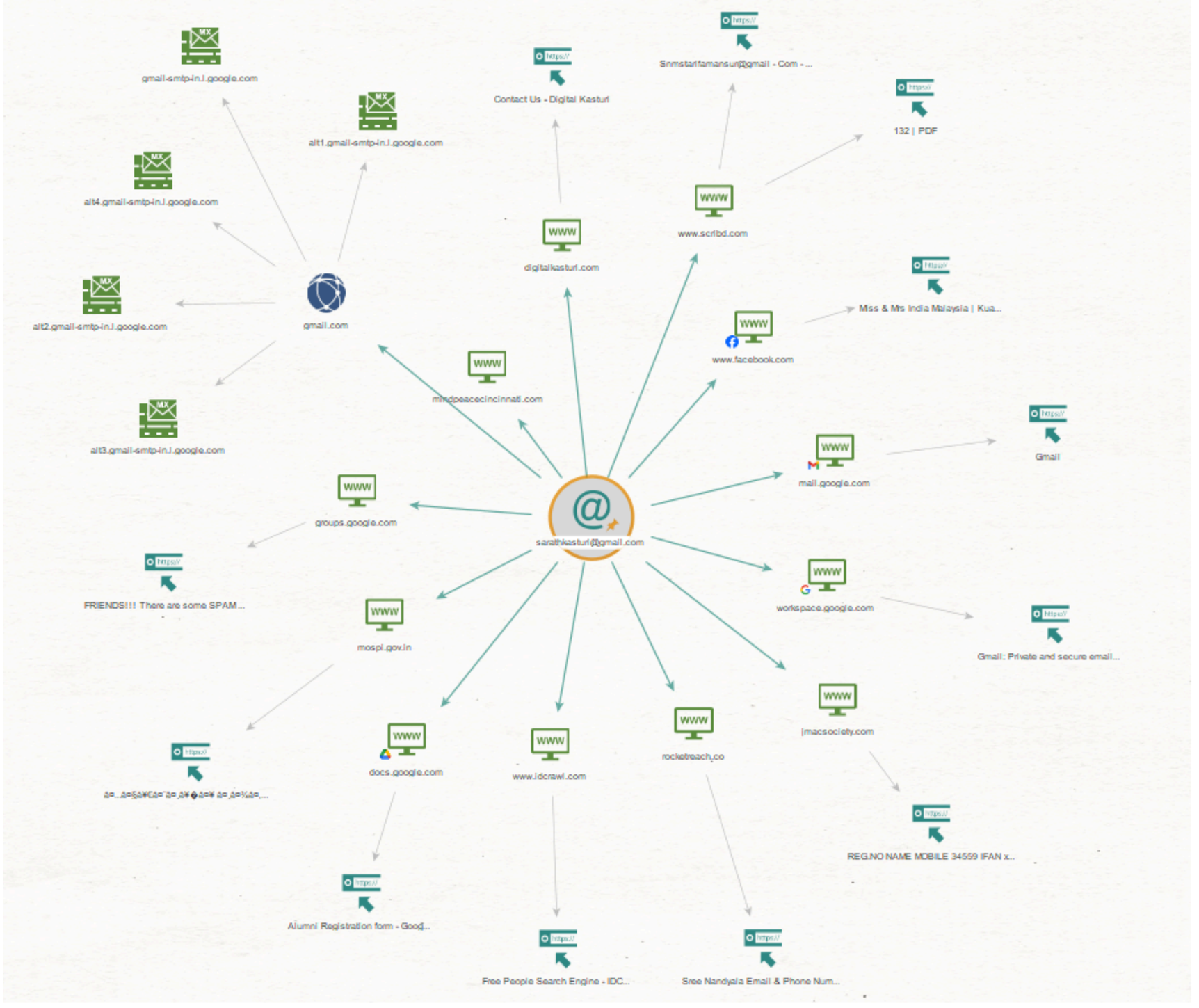
- Added seller email from the scraped dataset.
- Used transforms like “To Domain,” “To MX Record,” “To Entities”.
- Discovered related domains and infrastructure patterns.

Insights Gained:

- Identified sellers using the same domain/email host.
- Found repeated infrastructure footprints (same IP, domain, etc.).
- Helped link multiple listings possibly operated by the same actor.

Use in Investigation:

✓ Digital link analysis
✓ Network visualization
✓ Cross-checking seller identities


gmail-smtp-in.l.google.com
alt1.gmail-smtp-in.l.google.com
alt4.gmail-smtp-in.l.google.com
alt2.gmail-smtp-in.l.google.com
alt3.gmail-smtp-in.l.google.com
gmail.com
Contact Us - Digital Kasturi
digitalkasturi.com
Snmstarfamansur@gmail - Com - ...
132 | PDF
www.scribd.com
Miss & Mrs India Malaysia | Kua...
www.facebook.com
mindpeacecincinnati.com
groups.google.com
sarathkasturi@gmail.com
mail.google.com
Gmail
FRIENDS!!! There are some SPAM ...
mospi.gov.in
workspace.google.com
Gmail: Private and secure email ...
docs.google.com
www.idcrawl.com
rocketreach.co
imacsociety.com
REG.NO NAME MOBILE 34559 IFAN x...
Alumni Registration form - Good...
Free People Search Engine - IDC...
Sree Nandyala Email & Phone Num...

NumVerify

Steps Performed:

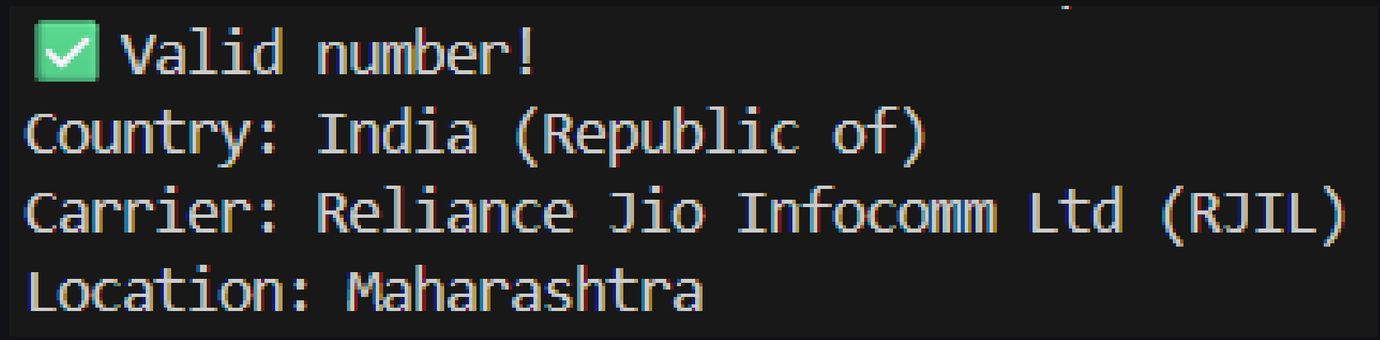
- Queried seller phone numbers using Numverify API.
- Collected metadata: location, carrier, line type (mobile/VoIP).
- Checked for reused or suspicious VoIP numbers.

Insights Gained:

- Flagged repeated or VoIP-based numbers.
- Detected mismatched country–region codes.
- Strengthened pattern recognition of fake seller networks.

Use in Investigation:

✓ Phone verification
✓ Pattern detection
✓ Cross-linking reused contacts



Conclusion

- Built a full crime-investigation simulation using synthetic marketplace data
- Web scraping automated evidence collection from listings
- Flags & ML detected reused images, contacts, and suspicious listings
- Network visualization (Gephi) exposed hidden seller clusters & link patterns
- OSINT tools (Maltego, Numverify) enriched investigation with external intelligence
- Demonstrated how data science + OSINT can uncover fraud networks and illicit trade patterns
- All data was synthetic & ethical, proving the workflow is safe and reproducible