PCOS Symptom Network Analysis

Technical Report

Lab Course: Network Science and Telematics

Submitted by: Iroda Ulmasboeva

Contact: ulmasboevairoda@gmail.com

Abstract

This report presents a hands-on application of network science to analyze symptom relationships in Polycystic Ovary Syndrome (PCOS). Using real-world clinical data and a modular Python-based pipeline, I followed a workflow that included data cleaning, transformation, network construction, and visualization. The primary goal was to uncover patterns of symptom co-occurrence in PCOS that might point to underlying clinical subtypes. All results reflect a specific configuration, though the pipeline is designed to be flexible.

Introduction

PCOS is a complex condition that influences multiple aspects of health—hormonal balance, metabolism, and reproductive function. Because symptoms vary significantly across patients, diagnosis and treatment are often inconsistent. In this project, I applied network science methods to understand how different symptoms connect, potentially helping to identify clusters of related symptoms or subtypes of the syndrome.

Methodology

1. Data Acquisition

The dataset used in this project was sourced from Kaggle: PCOS: A Guide to Practical Machine Learning. The specific file, `PCOS_data_without_infertility.xlsx`, focuses on clinical features unrelated to infertility.

2. Configuration

Project parameters were managed via `config.yaml`. This file defined:

- File paths and data locations
- Cleaning thresholds (e.g., removing columns with 50% missing data)
- Number of bins for discretizing numeric features
- The minimum edge weight for symptom co-occurrence (set to 150 for this analysis)

These parameters are fully customizable.

3. Data Cleaning

Using `data_cleaning.py`, I prepared the data by:

- Dropping features with excessive missing values
- Keeping the PCOS diagnosis column as a key label
- Binning numerical variables into three levels
- Removing identifier columns
- Creating binary flags for each binned feature

Outputs:

- cleaned_pcos_data.csv
- binary_transformed_pcos_data.csv
- patients_with_pcos.csv
- patients_without_pcos.csv
- binning_metadata.json

4. Symptom Co-occurrence Matrix

With `symptom_coocurence.py`, I generated pairwise co-occurrence counts of symptoms separately for PCOS and non-PCOS patients. The high threshold is set to include only symptom pairs appearing together at least 150 times to reduce the runtime.

5. Network Analysis

The actual networks were built using `network_utils.py`:

- Each node represents a binned symptom
- Edges connect symptoms that frequently co-occur
- Basic metrics: node degree and network density

Key stats:

- PCOS: 12 nodes, 58 edges, density = 0.8788
- Non-PCOS: 32 nodes, 371 edges, density = 0.7480

For community detection, I used the Louvain algorithm, which allowed me to identify potential subgroups of related symptoms.

6. Visualization

To make the results easier to interpret, I used `symptom_network_visuals.py` to create:

- Static images of each network
- Heatmaps to illustrate strong co-occurrence relationships
- Diagrams showing detected communities
- An interactive HTML network for exploration in the browser

Results

PCOS Network: A compact graph with 12 nodes and 58 edges. The most central symptoms were:

- amh(ng/ml) (0-22) — Anti-Müllerian Hormone
- fsh(miu/ml) (0-1684) — Follicle-Stimulating Hormone
- fsh/lh (0-458) — FSH to LH ratio
- prg(ng/ml) (0-28) — Progesterone

Non-PCOS Network: Much larger, with 32 nodes and 371 edges. It captured a broader range of clinical features, such as metabolic indicators and physical traits.

Community Detection:

- The PCOS network revealed tight clusters of hormonal and reproductive symptoms
- Non-PCOS networks were more diffuse, indicating less symptom clustering

Discussion

The networks show clear structural differences between PCOS and non-PCOS groups. PCOS networks had denser connections among a smaller number of symptoms—particularly hormone-related ones—while the non-PCOS group presented more varied and less tightly linked features. These findings support the idea that PCOS is characterized by a core cluster of interacting symptoms.

These patterns are influenced by configuration choices. For instance, lowering the co-occurrence threshold would increase network density but might also introduce noise. I experimented with several settings before settling on 150, which produced networks that were dense enough to reveal structure without becoming visually overwhelming.

Ideas for improvement:

- Explore how symptom connections change over time (dynamic networks)
- Try alternative centrality measures (e.g., betweenness, eigenvector)
- Compare detected communities with medical classifications of PCOS subtypes

Conclusion

This project combines clinical data with network analysis to reveal how symptoms cluster in PCOS. By turning raw data into a network structure, I was able to observe key patterns that might otherwise remain hidden. The flexibility of the approach means it can be adapted for other multifactorial conditions—and improved further with additional data sources or advanced metrics.

Acknowledgments

Prepared as part of the **Network Science and Telematics** lab course. Special thanks to my instructors: Sneha Mohanty and Prof. Dr. Christian Schindelhauer.

Project Repository

The full source code, data processing scripts, and visualizations can be found on GitHub:
https://github.com/iroooooda/PCOS_Symptom_Network_Analysis

Figures

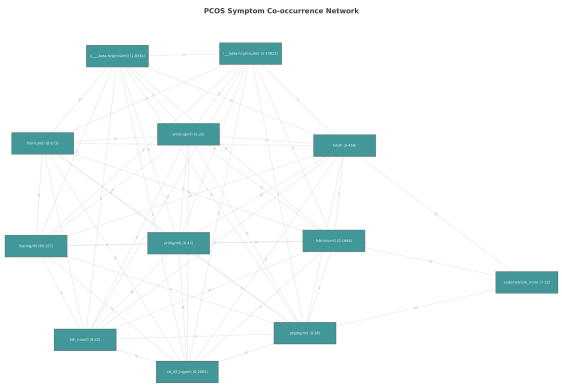

Figure 1: PCOS symptom co-occurrence network

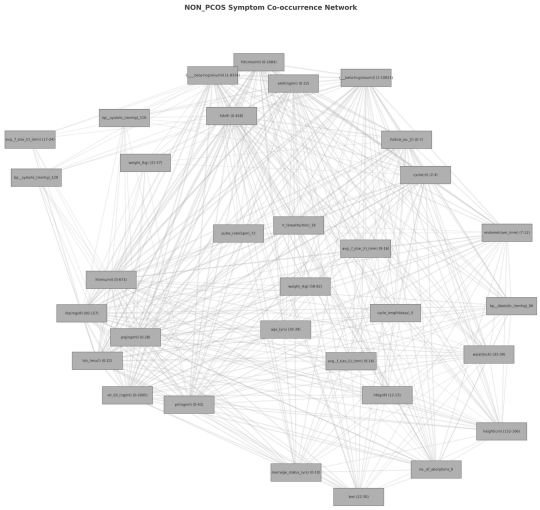

Figure 2: Non-PCOS symptom co-occurrence network

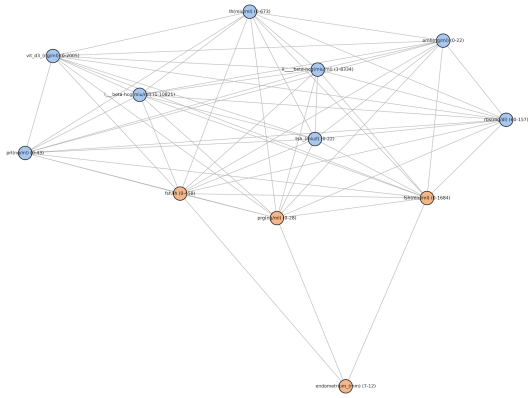

Figure 3: Louvain communities in PCOS symptom network

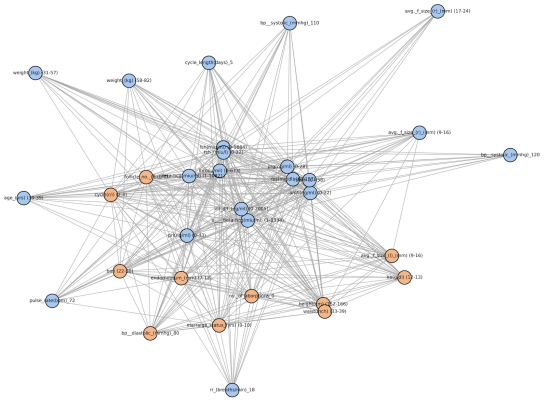

Figure 4: Louvain communities in non-PCOS symptom network

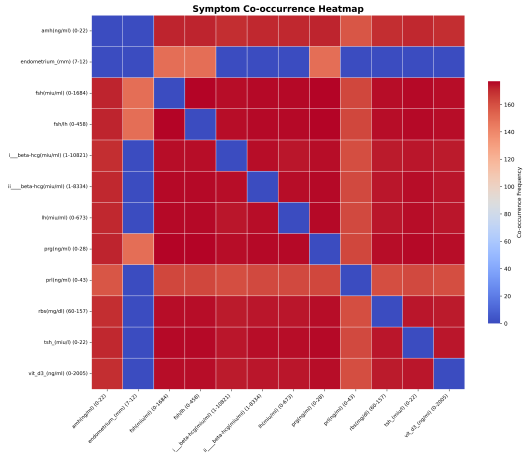

Figure 5: PCOS symptom co-occurrence heatmap

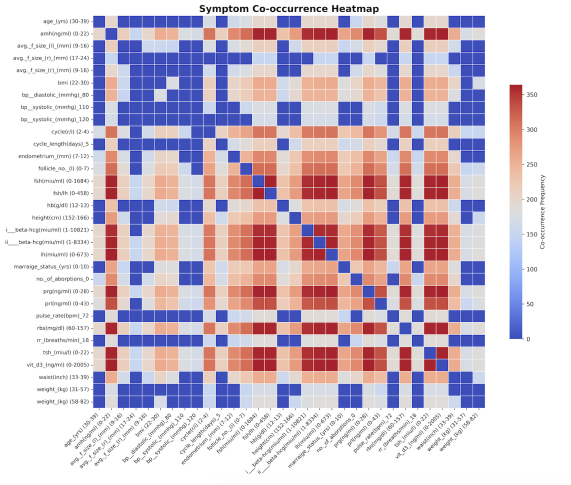

Figure 6: Non-PCOS symptom co-occurrence heatmap

References

- Kaggle. *PCOS: A Guide to Practical Machine Learning*. https://www.kaggle.com/code/psvenom/pcos-a-guide-to-practical-machine-learning-93-5
- Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). *Fast unfolding of communities in large networks*. Journal of Statistical Mechanics, 2008(10), P10008.