Predicting Membership in Healthy-Lifestyle Communities Using Network Science

research project presentation

Zuzanna Bąk
zuzanna.bak@temple.edu
MS Computational Data Science student

CIS 5524: ANALYSIS AND MODELING OF SOCIAL AND INFORMATION NETWORKS
Spring 2025

Agenda

1. Objective & Significance
2. Background
3. Proposed Approach
4. Data Description
5. Evaluation
6. Preliminary Results
7. Moving Forward
8. Discussion & Conclusions

TEMPLE UNIVERSITY

Objective

“Investigate the effect of local network influence on the adoption of healthy-lifestyle communities (subreddits) and predict which subreddits will ‘go healthy’ over time.”

Significance

Health Promotion: Identifying how health-related topics spread can help in designing targeted interventions or recommendations.

Network Science Contribution: Provides empirical evidence on how local connections (neighbors) correlate with the spread of specific interests.

Practical Use: Could be applied to recommendation systems or community detection for marketing, content moderation, or public health campaigns.

Homegrown Herb Garden
“THOSE WHO THINK THEY HAVE NO TIME FOR HEALTHY EATING WILL SOONER OR LATER HAVE TO FIND TIME FOR ILLNESS.”
– EDWARD STANLEY

Background

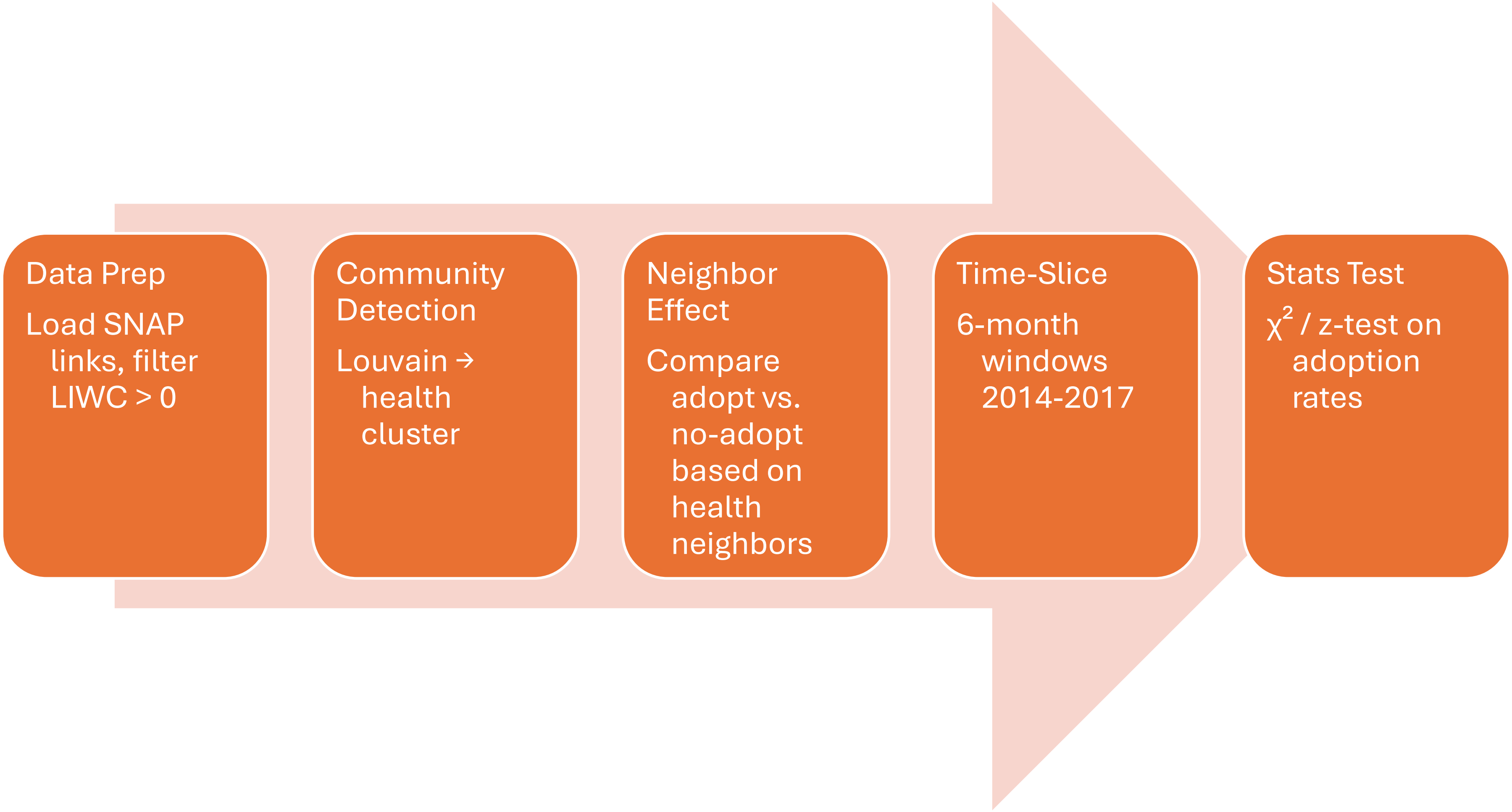
Network Science Fundamentals:

- *Small-world networks* and *community structure* (Barabási, Watts & Strogatz).
- *Local influence* in adoption—nodes with neighbors in a certain "state" are more likely to adopt.

Prior Work:

- Influence maximization (Kempe et al.) and link prediction (Liben-Nowell & Kleinberg) mostly show how ideas/behaviors diffuse.
- However, few studies specifically address **healthy-lifestyle** adoption in online forums (like Reddit).

Proposed Approach



Data Prep

Load SNAP links, filter LIWC > 0

Community Detection

Louvain → health cluster

Neighbor Effect

Compare adopt vs. no-adopt based on health neighbors

Time-Slice

6-month windows 2014-2017

Stats Test

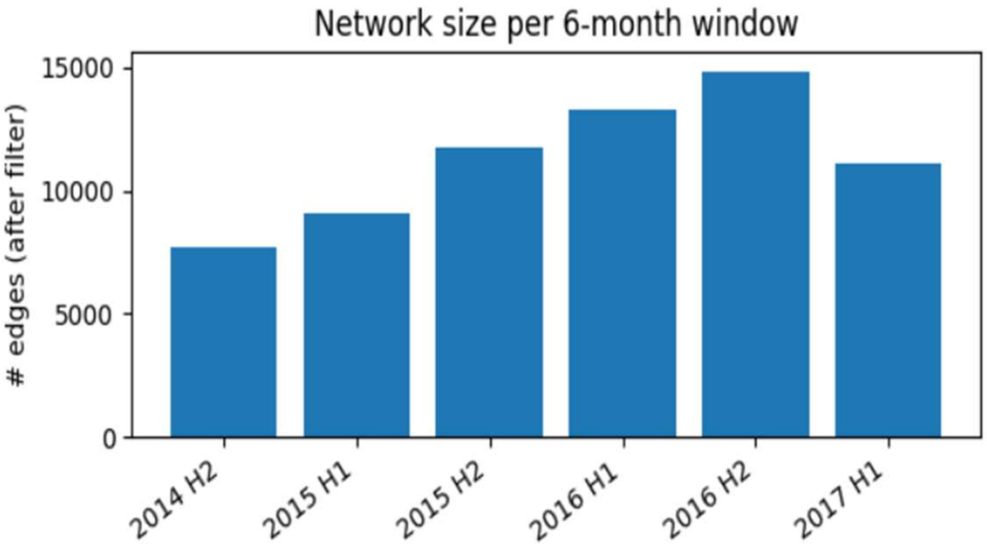
χ^2 / z-test on adoption rates

Data Description

Source: Stanford SNAP – Reddit Hyperlinks
Time span: Jan 2014 – Jun 2017
Raw edges: 850 k **Raw subreddits:** 55 k
Kept after filter (Body > 0 ∨ Health > 0):

- **Edges:** 286 k (34 %)
- **Subreddits:** 16 k (29 %)

6-Month Window	Nodes after filter	Edges after filter
2014 H2	4 119	7 667
2015 H1	4 876	9 042
2015 H2	6 141	11 755
2016 H1	6 960	13 310
2016 H2	7 599	14 857
2017 H1	6 117	11 094



Evaluation Plan

Define “Healthy-Lifestyle”	• (LIWC_Body + LIWC_Health) / 2 >= 0.01.
Adoption	• A subreddit not healthy in one window but labeled healthy in the next.
Neighbor Effect	• Probability of adoption for subreddits with ≥1 healthy neighbor vs. 0 healthy neighbors.
Metrics	• Compare probabilities; run chi-square / proportions z-test for significance.

Preliminary Results

- **Key Findings**
  - *With Healthy Neighbor:* ~8–10% adopt
  - *No Healthy Neighbor:* ~3–4% adopt
  - p-values < 10^-7 (highly significant difference)
- **Interpretation:** Subreddits with a healthy-lifestyle neighbor are ~2–3 times more likely to become healthy-lifestyle in the next 6-month window.

Window	T1 Range	T2 Range	T1 Graph (Nodes / Edges)	T2 Graph (Nodes / Edges)	Healthy T1 → T2 [New Adopters]	Adoption Probability (With vs. Without Neighbor)	Chi-Square (p-value)
1	2014-01-01 to 2014-07-01	2014-07-01 to 2015-01-01	4119 / 7667	4876 / 9042	601 → 714 [576]	0.0998 vs 0.0398	31.3038 (2.21e-08)

Preliminary Results

- **Key Findings**
  - *With Healthy Neighbor:* ~8–10% adopt
  - *No Healthy Neighbor:* ~3–4% adopt
  - p-values < 10^-7 (highly significant difference)
- **Interpretation:** Subreddits with a healthy-lifestyle neighbor are ~2–3 times more likely to become healthy-lifestyle in the next 6-month window.

Window	T1 Range	T2 Range	T1 Graph (Nodes / Edges)	T2 Graph (Nodes / Edges)	Healthy T1 → T2 [New Adopters]	Adoption Probability (With vs. Without Neighbor)	Chi-Square (p-value)
1	2014-01-01 to 2014-07-01	2014-07-01 to 2015-01-01	4119 / 7667	4876 / 9042	601 → 714 [576]	0.0998 vs 0.0398	31.3038 (2.21e-08)
2	2014-07-01 to 2015-01-01	2015-01-01 to 2015-07-01	4876 / 9042	6141 / 11755	714 → 827 [687]	0.0980 vs 0.0435	34.3822 (4.53e-09)

Preliminary Results

- **Key Findings**
  - *With Healthy Neighbor:* ~8–10% adopt
  - *No Healthy Neighbor:* ~3–4% adopt
  - p-values < 10^-7 (highly significant difference)
- **Interpretation:** Subreddits with a healthy-lifestyle neighbor are ~2–3 times more likely to become healthy-lifestyle in the next 6-month window.

Window	T1 Range	T2 Range	T1 Graph (Nodes / Edges)	T2 Graph (Nodes / Edges)	Healthy T1 → T2 [New Adopters]	Adoption Probability (With vs. Without Neighbor)	Chi-Square (p-value)
1	2014-01-01 to 2014-07-01	2014-07-01 to 2015-01-01	4119 / 7667	4876 / 9042	601 → 714 [576]	0.0998 vs 0.0398	31.3038 (2.21e-08)
2	2014-07-01 to 2015-01-01	2015-01-01 to 2015-07-01	4876 / 9042	6141 / 11755	714 → 827 [687]	0.0980 vs 0.0435	34.3822 (4.53e-09)
3	2015-01-01 to 2015-07-01	2015-07-01 to 2016-01-01	6141 / 11755	6960 / 13310	827 → 865 [719]	0.0772 vs 0.0355	28.2926 (1.04e-07)
4	2015-07-01 to 2016-01-01	2016-01-01 to 2016-07-01	6960 / 13310	7340 / 14857	865 → 1026 [843]	0.0803 vs 0.0398	29.3069 (6.18e-08)
5	2016-01-01 to 2016-07-01	2016-07-01 to 2017-01-01	7340 / 14857	7599 / 14025	1026 → 1079 [876]	0.1072 vs 0.0356	93.3189 (4.45e-22)
6	2016-07-01 to 2017-01-01	2017-01-01 to 2017-07-01	7599 / 14025	6117 / 11094	1079 → 850 [670]	0.1060 vs 0.0282	129.5257 (5.20e-30)

Moving Forward → Final Report (due May 1, 5:30 pm)

TEMPLE UNIVERSITY

1. Refine "Healthy-Lifestyle" Label

- Test alternative LIWC thresholds (0.005 – 0.02)
- Add text-embedding check (SBERT) to capture fitness keywords not in LIWC

2. Deeper Network Analysis

- Run logistic regression: health-neighbor + degree + activity
- Compute centrality (betweenness, eigenvector) as additional predictors
- Repeat adoption test on quarterly windows for robustness

3. Causality vs. Homophily

- Propensity-score matching: balance on prior health language & degree
- Compare matched pairs' adoption rates → report ATT & confidence interval

4. Final-Report Package (PDF + supplementals)

- Full results tables & code repo link
- Limitations + future-work section (2 paragraphs)
- APA-formatted references

(Everything above scheduled; no additional data collection needed.)

Discussion & Conclusions

TEMPLE UNIVERSITY

Discussion:

- The consistent neighbor effect suggests local exposure drives adoption of health topics.
- Statistically significant across all windows, indicating the phenomenon is robust over time.

Limitations:

- Observational data → can't prove strict causality.
- LIWC thresholds may not perfectly capture health-related content.

Next Steps:

- Incorporate advanced community detection (e.g., Louvain).
- Possibly compare other topics (e.g., diet vs. fitness sub-communities) to see if patterns differ.

Conclusion:

- The project so far supports the hypothesis that local connectivity **strongly** correlates with healthy-lifestyle adoption on Reddit.

Thank you

Any questions?

Zuzanna Bąk
zuzanna.bak@temple.edu
MS Computational Data Science student

References

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science, 329*(5996), 1194–1197.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America, 113*(3), 554–559.

Forbes / McCarthy, N. (2020, August 21). *Report: Misinformation on Facebook poses a major threat to public health [Infographic]*. Retrieved March 24th , 2025 from https://www.forbes.com/sites/niallmccarthy/2020/08/21/report-misinformation-on-facebook-poses-a-major-threat-to-public-health-infographic/

KGUN9. (n.d.). *Different types of misinformation and how to identify it*. Retrieved March 24th , 2025 from https://www.kgun9.com/news-literacy-project/different-types-of-misinformation-and-how-to-identify-it

MDPI – Information, 15(1), Article 60. (2023). *Mapping the Landscape of Misinformation Detection: A Bibliometric Approach*. Retrieved March 24th , 2025 from https://www.mdpi.com/2078-2489/15/1/60

ResearchGate – Independent Cascade Model. (n.d.). *Illustration of the independent cascade model: The decomposition diagrams of four time steps*... Retrieved March 24th , 2025 from https://www.researchgate.net/figure/Illustration-of-the-independent-cascade-model-The-decomposition-diagrams-of-four-time_fig5_369791091

ResearchGate – SIR Model in Scale-Free Network. (n.d.). *Example of an epidemic situation by applying SIR model to scale-free network*. Retrieved March 24th , 2025 from https://www.researchgate.net/figure/Example-of-an-epidemic-situation-by-applying-SIR-model-to-scale-free-network-Snapshot-of_fig3_317711741

ResearchGate – Schematic View of Network Concepts. (n.d.). Retrieved March 24th , 2025 from https://www.researchgate.net/figure/Schematic-view-of-network-concepts-nodes-edges-hubs-centrality-and-connectivity_fig1_260197221

ResearchGate – The spreading of misinformation online. (2016). Retrieved March 24th , 2025 from https://www.researchgate.net/publication/289263634_The_spreading_of_misinformation_online

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.

YourDictionary. (n.d.). *Misinformation vs. disinformation: Compare*. Retrieved March 24th , 2025 from https://www.yourdictionary.com/articles/misinformation-disinformation-compare