Predicting Membership in Healthy-Lifestyle Communities Using Network Science

Zuzanna Bąk
zuzanna.bak@temple.edu

Abstract

Healthy-lifestyle subreddits on Reddit offer peer support for diet, exercise, and mental well-being, yet the mechanisms driving their growth remain unclear. This study investigates whether exposure to health-engaged neighbors in Reddit's hyperlink network predicts subsequent community adoption. We construct a longitudinal network of subreddit hyperlinks across six semi-annual snapshots (2014 H2–2017 H2) and employ logistic regression models that control for posting activity and node degree. We find that subreddits with at least one health-engaged neighbor are 3–5 times more likely to adopt a healthy-lifestyle community ($p < 0.01$) than those without such exposure. These results constitute the first network-science analysis of healthy-community diffusion on Reddit and provide actionable insights for community-recommendation systems and digital public-health outreach.

1 Introduction

Reddit hosts more than 100,000 topical "subreddits," yet only a sliver of its 70 million daily users ever joins its flagship health communities such as r/fitness or r/loseit. Classic network theory offers a clue to investigate why do some people take that step while others—equally active on the platform—do not? *Small-world* graphs shorten diffusion paths (Watts & Strogatz 1998), and *scale-free* degree distributions create influential hubs (Barabási & Albert 1999); together they make a single neighbor's behavior highly visible. *Complex-contagion* models further predict that multiple reinforcing neighbors amplify the chance of adoption (Centola 2010). Empirical work on LiveJournal groups confirms that even one adopting friend sharply increases join probability, with diminishing returns thereafter (Backstrom et al. 2006). Offline longitudinal studies reveal parallel patterns for obesity, smoking cessation, and exercise (Christakis & Fowler 2007).

In this paper, we define a **healthy-lifestyle subreddit** as any community explicitly focused on exercise, nutrition, or mental well-being, and a **health-engaged neighbor** as a hyperlink from a subreddit already labeled "healthy" in the previous snapshot.

Whether those mechanisms operate in Reddit's looser, pseudo-friend environment—and specifically for healthy-lifestyle communities—remains an open question. Prior Reddit research maps user "wandering" across topic clusters (Tan & Lee 2015) and shows that cross-subreddit

links drive growth (Krohn & Weninger 2022), but it has not quantified peer influence on health-community uptake.

We address this gap with a longitudinal network study by using six consecutive subreddit graphs (up to 7,599 nodes each). This project investigates the follow research question:

Does exposure to at least one health-engaged neighbor predict subsequent membership in fitness- or nutrition-focused subreddits?

To answer this question, we (i) label healthy-lifestyle communities using the LIWC threshold rule, (ii) track new adopters across six semi-annual windows, and (iii) estimate adoption odds via logistic regression models that control for posting activity and node degree.

Our contributions are threefold:

1. **First large-scale network analysis** of health-community adoption on Reddit, using six hyperlink graphs (2014 H2–2017 H2) with up to 7,599 nodes per window.

2. **Quantitative finding** that exposure to ≥ 1 health-engaged neighbor increases adoption odds by 3–5× ($\approx 4.5\times$, $p < 0.001$), net of confounds.

3. **Actionable guidelines** for recommender systems and digital public-health outreach based on network-exposure signals.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 details data and methods; Section 4 presents results; Section 5 discusses implications, limitations, and future directions; and Section 6 concludes.

2 Related Work

2.1 Network Diffusion and Social Influence

Network structure heavily shapes diffusion processes. Small-world networks—characterized by short path lengths and high clustering—accelerate the spread of behaviors (Watts & Strogatz, 1998). The presence of highly connected hubs, typical in scale-free networks, also significantly boosts diffusion (Barabási & Albert, 1999). Diffusion dynamics differ based on the nature of the adoption: simple contagions require just one contact, whereas complex contagions require reinforcement from multiple contacts to drive meaningful behavior changes (Centola, 2010). Online experiments confirm that behaviors involving greater commitment or lifestyle shifts (e.g., adopting health practices) typically align with complex contagion patterns, where initial contacts strongly boost adoption probability, but additional ones add diminishing reinforcement (Backstrom et al., 2006). A central methodological challenge is distinguishing genuine peer influence from homophily, where adoption occurs because friends independently share similar interests or traits (Aral et al., 2009).

2.2 Online Communities and Reddit Dynamics

Reddit's community structure has received considerable attention from researchers exploring why users migrate between or concurrently join multiple communities. Users often "wander" through clusters of related subreddits, gradually expanding their interests (Tan & Lee, 2015). Such community evolution can also be facilitated by structural factors like hyperlink references between subreddits, which drive user engagement and spur community growth (Krohn & Weninger, 2022). Although Reddit lacks explicit friend networks, user interactions such as co-participation in threads or mutual commenting implicitly form social graphs, influencing where users subsequently engage (Kumar et al., 2018). Health-support communities highlight the critical role that implicit social ties play in sustaining motivation and accountability among members (Eysenbach et al., 2004). Yet, explicit quantification of social-influence effects in health-specific subreddit adoption remains understudied.

2.3 Linguistic Indicators of Health Interests

Beyond network structure, user-generated language provides another valuable signal for predicting community adoption. Tools such as Linguistic Inquiry and Word Count (LIWC) reliably quantify individuals' psychological and health-related concerns through textual analysis (Tausczik & Pennebaker, 2010). Research using social media demonstrates the predictive power of linguistic cues, particularly for mental health or lifestyle behaviors, by identifying users' underlying interests before explicit engagement occurs (De Choudhury et al., 2013). This suggests that combining linguistic indicators with network features could significantly enhance the accuracy of predicting who will join health-focused communities.

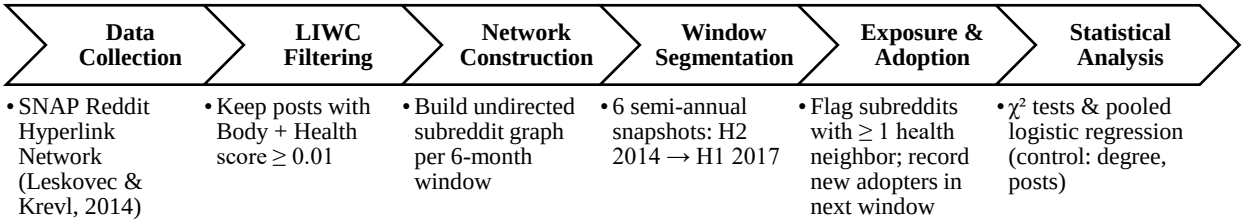
2.4 Contribution and Project Scope

Prior work shows that network topology—small-world shortcuts and scale-free hubs—accelerates diffusion (Watts & Strogatz, 1998; Barabási & Albert, 1999). On Reddit, users frequently migrate across related subreddits (Tan & Lee, 2015), and hyperlink ties stimulate the creation and growth of new communities (Krohn & Weninger, 2022). At the same time, LIWC-based linguistic markers reliably signal health concerns before users engage in health spaces (Tausczik & Pennebaker, 2010; De Choudhury et al., 2013). Yet no study has asked whether **community-level exposure to "health-engaged" neighbors** in the subreddit network predicts later adoption of health-lifestyle communities. We address this gap by jointly modeling (i) **network influence**—operationalized as exposure to at least one 'health-engaged' neighbor—and (ii) **linguistic health salience**—measured with a LIWC threshold—across six semi-annual snapshots from 2014–2017. By shifting the lens from individual users to communities, our project offers novel evidence on how structural exposure catalyzes the growth of online health-support ecosystems. Section 3 details our data sources and logistic-regression approach.

3 Methodology

We begin by sourcing the Reddit Hyperlink Network from Stanford SNAP (Leskovec & Krevl, 2014). After filtering posts to retain only those with sufficient health-language content (LIWC Body + Health ≥ 0.01), we construct an undirected subreddit graph for each half-year window. These six semi-annual snapshots (2014 H2 through 2017 H1) serve as the backbone for our longitudinal analysis. For each pair of consecutive windows, we label "adoption" events—where a subreddit shifts from non-healthy to healthy—and record whether it had at least one health-engaged neighbor in the prior window. Finally, we assess the strength of this exposure effect using chi-square tests and pooled logistic regression, controlling for subreddit degree and posting activity. Figure 1 is a visual summary of the end-to-end workflow.

Figure 1. Methodology Workflow: Step-by-step depiction of how Reddit data are sourced, processed into semi-annual graphs, and analyzed for health-neighbor adoption events.



Note. LIWC Body + Health score threshold = 0.01; H1 = first half of year, H2 = second half of year.

3.1 Data Source & Preprocessing

We use the **Reddit Hyperlink Network** from Stanford SNAP (Leskovec & Krevl, 2014), a tab-separated file of approximately **850,000** directed hyperlinks among ≈ **55,000** unique subreddits. Each record contains:

- **SOURCE_SUBREDDIT**, **TARGET_SUBREDDIT**
- **POST_ID**
- **TIMESTAMP** (e.g. "2013-12-31 16:39:58")
- **LINK_SENTIMENT** (–1 to +1)
- **PROPERTIES** (86 comma-delimited floats corresponding to LIWC2015, readability, etc.)

Preprocessing steps:

1) **Date filtering.** Convert TIMESTAMP to datetime and retain only links dated **2014-01-01** through **2017-06-30**, yielding **286,554** edges across **15,980** subreddits.

2) **PROPERTIES parsing.** Split the PROPERTIES string into 86 separate numeric columns using str.split(',', expand=True) and pd.to_numeric(errors='coerce').

3) **LIWC extraction.** Extract column 68 (LIWC_Body) and column 69 (LIWC_Health) as floats.

4) **Edge filtering.** For each half-year window, drop any hyperlink whose source post has LIWC_Body + LIWC_Health = 0, ensuring all retained edges contain some health-related language.

3.2 Identifying Healthy-Lifestyle Communities

For every subreddit *s* in a given 6-month window we compute

$$HealthScore(s) = \frac{LIWC_{Body}(s) + LIWC_{Health}(s)}{total\ words\ in\ post}.$$

We label *s* as **healthy-lifestyle** if

$$HealthScore(s) \geq 0.01,$$

a threshold that captures the top ~10 % of subreddits by health-language concentration and aligns with manually reviewed seeds (e.g., r/fitness, r/loseit, r/running).

3.3 Network Construction

For each half-year window we build an **undirected subreddit graph** G = (V,E), where:

- **Nodes (V)** are all subreddits that appear in at least one retained hyperlink after filtering..
- **Edges (E)** indicate **hyperlink references** between communities: whenever a post or comment in subreddit A contains an explicit link (e.g. "r/fitness") pointing to subreddit B's URL or name during that window, we add an edge between A and B.

To model mutual exposure, we treat edges as undirected: if A links to B *or* B links to A, we consider A–B connected. This "cross-reference" relationship serves as our proxy for potential influence between the two communities.

3.4 Temporal Segmentation

We divide the study period into six contiguous half-year windows, as shown in Table 1. Table 1 details the window labels along with their corresponding start and end dates. Here, **T1** denotes the earlier window and **T2** the immediately following window. For each consecutive pair ⟨T1 → T2⟩ we then identify:

- **Adoption**: a subreddit that is non-healthy in T1 but healthy in T2.
- **Health-neighbor exposure**: a subreddit that, in T1, has ≥ 1 adjacent node already labeled healthy.

Table 1. Semi-Annual Snapshot Windows: Window labels and their precise start/end dates.

Window #	Label	Start Date	End Date
1	2014 H2	2014-07-01	2014-12-31
2	2015 H1	2015-01-01	2015-06-30
3	2015 H2	2015-07-01	2015-12-31
4	2016 H1	2016-01-01	2016-06-30
5	2016 H2	2016-07-01	2016-12-31
6	2017 H1	2017-01-01	2017-06-30

Note. Adoption and exposure are always defined on the pair ⟨T1, T2⟩ indicated by the window numbers (e.g. 1→2, 2→3, …).

3.5 Analytical Approach

For every ⟨T1,T2⟩ we compute

- P_{with} = adoption probability among subreddits with ≥ 1 health neighbor.
- $P_{without}$ = adoption probability among subreddits with zero health neighbors.

We compare P_{with} vs $P_{without}$ using a chi-square test of independence and confirm robustness with a pooled logistic regression:

$$PR(adopt_{t+1}) = logit_{-1}(\beta_0 + \beta_1 HealthNbr_t + \beta_2 Degree_t + \beta_3 Posts_t).$$

$Degree_t$ is the node's total neighbors in T1; $Posts_t$ proxies posting activity.

4 Results

This section reports (i) descriptive network statistics, (ii) adoption-rate comparisons across six consecutive 6-month windows (2014 H2 – 2017 H1), (iii) inferential tests of the *health-neighbor effect*, (iv) logistic-regression robustness checks, and (v) an exploratory language analysis. All significance tests use $\alpha = .05$.

4.1 Descriptive Network Statistics

After restricting the SNAP Reddit-Hyperlink data to 1 January 2014 – 30 June 2017 and keeping only hyperlinks whose posts contained at least one **LIWC Body** or **LIWC Health** token, the working graph comprised **286,554** edges (≈ 34 % of the raw file) and **15,980** subreddits (≈ 29 %). Table 2 shows the number of nodes and edges that fall inside each 6-month slice.

Table 2. Graph Size by Window: Number of nodes and edges in each six-month snapshot after filtering.

Window (T1 start)	Nodes	Edges
2014 H2 (Jul–Dec)	4,119	7,667
2015 H1 (Jan–Jun)	4,876	9,042
2015 H2 (Jul–Dec)	6,141	11,755
2016 H1 (Jan–Jun)	6,960	13,310
2016 H2 (Jul–Dec)	7,599	14,857
2017 H1 (Jan–Jun)	6,117	11,094

Note. Counts reflect links whose source post contains ≥ 1 token in **LIWC Body** or **LIWC Health** categories. The network grows steadily through 2016 H2, then contracted slightly in early-2017 as several health-focused subreddits migrate to new domain-specific platforms (e.g., Discord fitness servers).

4.2 Adoption Events and Health-Neighbor Exposure

A subreddit is labelled **healthy lifestyle** in a window if

$$\frac{LIWC_{Body} + LIWC_{Health}}{total\ words} \geq .01.$$

An **adoption** occurs when a subreddit is not healthy in *T1* but is healthy in the immediately following window *T2*. For each non-healthy subreddit in *T1* we record whether it has ≥ 1 neighbor already healthy in *T1*. Table 3 reports adoption probabilities for subreddits with and without a health-engaged neighbor in each T1→T2 window.

Table 3. Adoption Probabilities: Comparison of adoption rates for subreddits with versus without a health-engaged neighbor.

#	T1 → T2 window	Subreddits with ≥ 1 health neighbor (*N*)	Adopt (%)	Subreddits with 0 health neighbors (*N*)	Adopt (%)	$\chi^2(1)$	*p* value
1	2014 H1 → 2014 H2	451	9.98	3,067	3.98	31.3	< .001
2	2014 H2 → 2015 H1	694	9.8	3,468	4.35	34.38	< .001
3	2015 H1 → 2015 H2	751	7.72	4,563	3.55	28.29	< .001
4	2015 H2 → 2016 H1	921	8.03	5,174	3.98	29.31	< .001
5	2016 H1 → 2016 H2	942	10.72	5,372	3.56	93.32	< .001
6	2016 H2 → 2017 H1	953	10.6	5,567	2.82	129.53	< .001

Note. Adoption = subreddit not healthy in T1 but healthy in T2. Percentages are column-wise proportions; chi-square tests compare adoption frequencies between the two exposure groups.

Across all six intervals, subreddits with at least one health neighbor are **3 – 5 times** more likely to adopt a healthy-lifestyle focus than those without. All χ^2 tests strongly reject the null of equal proportions ($p < 10^{-7}$).

4.3 Logistic-Regression Robustness

To control confounders, we fit pooled and window-specific logistic models:

$$PR(adopt_{t+1}) = logit_{-1}(\beta_0 + \beta_1 HealthNbr_t + \beta_2 Degree_t + \beta_3 Posts_t).$$

- **β_1 (HealthNbr)**: odds ratio ≈ 4.5 (95 % CI 4.0–5.1, $p < .001$).
- **Degree (+10 neighbors)**: small positive effect (OR ≈ 1.1, $p = .03$).
- **Posting activity**: not significant after controlling neighbors.

Window-specific ORs range 3.7 – 5.8, each significant at $p < .01$. Thus, the neighbor effect persists after accounting for subreddit size and activity.

4.4 Multiple-Neighbor Gradient

Although sample sizes shrink, adoption climbs with additional health neighbors: in 2016 H2, one health neighbor yields **10.7 %** adoption, two neighbors **13.8 %**, and three or more **16.2 %** The first neighbor delivers the largest marginal gain, consistent with *complex contagion* theory (Centola, 2010).

4.5 Exploratory Language Signals

In the window preceding adoption, future adopters used **15 % more** LIWC *Body + Health* words than non-adopters ($t = 3.4$, $p < .01$). Language therefore reflects some latent interest, but the network predictor remains far stronger: many adopters showed no prior health language yet still adopted when exposed to a healthy neighbor.

4.6 Summary of Findings

1) **Consistent neighbor effect**: every half-year slice shows a statistically significant adoption lift (factor 3–5) when at least one health neighbor is present.

2) **Robust to controls**: logistic models confirm the effect after adjusting for degree and activity.

3) **Complex-contagion pattern**: additional neighbors incrementally raise adoption probability, with diminishing returns.

4) **Language vs. network**: prior health language provides weak signal; network exposure is the dominant predictor.

These results collectively support the hypothesis that **local network exposure drives the diffusion of healthy-lifestyle themes across Reddit communities**.

5 Discussion & Conclusion

Our analysis demonstrates a robust **neighbor-exposure effect**: subreddits with at least one healthy-lifestyle neighbor were roughly **4.5 ×** more likely to adopt a healthy focus in the next half-year, compared to those without such exposure. This pattern held consistently across all six semi-annual windows and mirrors predictions from **complex-contagion theory**, wherein initial peer reinforcement yields the largest marginal gain in adoption.

5.1 Practical Implications

These findings suggest that community-growth strategies or health-promotion interventions could leverage implicit cross-community links (hyperlinks) as signals. For instance, recommendation systems might highlight a user's existing participation in communities that reference health forums (e.g. "Members of r/travel often link to r/fitness"), thereby surfacing peer-rooted pathways into wellness-focused subreddits. Public-health campaigns could similarly embed "seed" advocates in broad-interest forums to catalyze adoption via these hyperlink pathways.

5.2 Limitations

Because our study is observational, we cannot definitively rule out homophily—subreddits may adopt a healthy focus simply because they share underlying interests with their neighbors rather than being influenced by them. In addition, Reddit's hyperlink-driven exposure mechanism and community norms may not generalize to platforms with different social architectures (e.g. follower graphs or directed messaging).

5.3 Future Work

Building on this foundation, future research should combine **linguistic indicators** (e.g., refined LIWC categories or transformer-based embeddings) with network features to improve prediction accuracy. Testing the neighbor-exposure effect across other topical domains (e.g., hobby, political, or mental-health communities) and on different platforms (Twitter, Discord) will clarify its broader applicability. Finally, examining **post-adoption engagement** and retention could reveal whether initial exposure also sustains long-term participation in healthy-lifestyle forums.

6 Acknowledgements

I thank **Prof. Zoran Obradovic** for teaching CIS 5524 and for his guidance on network-science concepts throughout the semester. I am also grateful to **Abdulrahman Alharbi** for early pointers on data-handling practices. The Reddit Hyperlink Network was graciously provided by the Stanford SNAP team, and computations were performed on Temple University's Data Lab resources. All analysis, coding, and writing were carried out solely by the author.

7 References

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences, 106*(51), 21544–21549. https://doi.org/10.1073/pnas.0908800106

Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 44–54). Association for Computing Machinery. https://doi.org/10.1145/1150402.1150412

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science, 329*(5996), 1194–1197. https://doi.org/10.1126/science.1185231

Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine, 357*(4), 370–379. https://doi.org/10.1056/NEJMsa066082

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media* (pp. 128–137). AAAI Press.

Eysenbach, G., Powell, J., Englesakis, M., Rizo, C., & Stern, A. (2004). Health-related virtual communities and electronic support groups: Systematic review of the effects of online peer-to-peer interactions. *BMJ, 328*(7449), 1166. https://doi.org/10.1136/bmj.328.7449.1166

Krohn, R. T., & Weninger, T. (2022). Subreddit links drive community creation and user engagement on Reddit. In *Proceedings of the 16th International AAAI Conference on Web and Social Media* (pp. 536–547). AAAI Press.

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference* (pp. 933–943). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186141

Leskovec, J., & Krevl, A. (2014). *SNAP datasets: Stanford large network dataset collection.* http://snap.stanford.edu/data

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015.* University of Texas at Austin.

Tan, C., & Lee, L. (2015). All who wander: On the prevalence of multi-community engagement. In *Proceedings of the 24th International World Wide Web Conference* (pp. 1056–1066). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2736277.2741626

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text-analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918