Assignment 2*

My Starbucks Idea

Sowmya Sathi
ssathi2@students.towson.edu

Abhishek Rangi
arangi1@students.towson.edu

Jaya Peda Vignesh Reddy Duggempudi
jduggem1@students.towson.edu

***Abstract*—We analyzed a stratified 1 500-idea sample—the 500 most-voted, 500 median-voted, and 500 least-voted suggestions—together with 17 328 comments from the MyStarbucks Idea platform. The data was modeled as four complementary network layers: a directed comment-flow graph, an undirected co-commenter graph, a bipartite user–idea graph, and a projected idea-similarity graph. Louvain community detection uncovered seven tightly knit user groups, with $\approx 82\%$ of comment traffic flowing inside rather than across communities (Q2). Role-mixing analysis shows only a mild assortativity (0.07) between staff"experts", contributors, and regular clients, indicating that official accounts interact almost as freely with customers as with each other (Q3). A ridge-regularised logistic model that uses comment volume and author betweenness (but not vote count) reaches an AUC of 0.90 on unseen ideas, demonstrating that network engagement signals strongly predict which suggestions rise to the top (Q4). These results illustrate how multi-layer network analytics, even on a representative sample, can guide crowd-sourcing strategy for corporate innovation.**

***Index Terms*—Complex networks, community detection, social media analytics, Louvain algorithm, logistic regression.**

I. Introduction

Digital crowdsourcing platforms have become a cornerstone of modern open-innovation strategies. By lowering participation barriers and making discussion threads public, these systems convert thousands of individual suggestions, comments and votes into a live marketplace of ideas. Firms benefit from faster product-innovation cycles and a direct line to loyal customers, while contributors gain influence, recognition and a sense of co-ownership. Starbucks adopted this model through *My Starbucks Idea* (MSI), an online venue that—at its peak—hosted more than two million votes and tens of thousands of proposals ranging from new beverages to sustainability practices.

Despite its promise, such a forum can easily bury good ideas in noise. Managers therefore face four unresolved questions:

1) *Data-to-network:* How should raw clicks, comments, and votes be translated into graph models that faithfully represent different kinds of interaction?
2) *Community structure:* What hidden groups of users exist, and how tightly—or loosely—are those groups connected to one another?
3) *Role ties:* Do distinct user roles (staff experts, heavy contributors, casual clients) form special connections that amplify or dampen influence?
4) *Idea success:* Which network signals predict whether a suggestion will flourish or die, and can those signals be captured in a practical model?

To address these questions we analyse a balanced, 1 500-idea slice of the MSI corpus (the 500 most-voted, 500 median-voted and 500 least-voted suggestions) together with 17,328 associated comments. Our study offers four contributions:

- Multi-layer graph model. We build four complementary networks—directed comment-flow, undirected co-commenter, user–idea bipartite, and idea-similarity—plus per-topic sub-graphs, capturing both user-to-user and user-to-idea relations.
- Community insights. Louvain detection on the co-commenter layer uncovers seven cohesive clusters, with $\approx 82\%$ of comment traffic flowing *within* rather than across communities.
- Role interaction analysis. A heuristic tagger separates *experts*, *contributors* and *clients*, revealing only mild assortativity ($r = 0.07$); staff accounts engage almost evenly across roles.
- Predictive success model. A ridge-regularised logistic regression—fed by comment volume and author betweenness—achieves an AUC $=0.90$ on a held-out test set, demonstrating that network engagement signals alone strongly predict idea prominence.

These findings illustrate how multi-layer network analytics, even on a representative sample, can turn raw crowdsourcing data into actionable guidance for corporate innovation teams.

II. Related Work

A. Crowdsourced Idea Platforms

Research on corporate idea marketplaces dates back to Dell's *Ideastorm* and Starbucks' own *My Starbucks Idea* (MSI) launch in 2008. User motivation and idea life-cycles on MSI were qualitatively examined by Oliveira et.al. [1], while Chen and Xie [2] compared adoption rates between Dell and Starbucks, highlighting managerial endorsement as a critical bottleneck. More recent work leverages network analytics: Vasan et al. [3] modelled Apple's *App Store Reviews* as a bipartite graph to rank feature requests; Xu et al. [4] studied the open-source GitHub "issue" tracker as a crowd-ideation venue. However, these studies typically focus on a single interaction layer (votes or comments), missing the multi-layer perspective.

B. Community Detection in Social Networks

Graph partitioning has become a standard tool for social-media mining. Newman and Girvan's modularity measure [5] remains the dominant quality function, with the Louvain algorithm [6] prized for its speed on million-node graphs. Applications range from Twitter hashtag tribes [7] to citation networks [8]. A strand of work examines comment forums: Hessel et.al. [9] used Louvain on Reddit threads to isolate "information silos," finding that intra-community reply fraction is a proxy for ideological homophily.

C. *Predicting Idea Success Online*

Logistic or survival models have been widely applied to *Kickstarter* [10], *Quora* [11], and *Stack Overflow* [12] to explain why some posts gain traction. Typical predictors include early vote velocity, author reputation, and textual sentiment. Chen et.al. [13] added network centrality of the asker to improve AUC by 6%. Few papers, however, test whether purely network-level features—in the absence of text or raw votes—can achieve comparable lift.

D. *Positioning of the Present Study*

Our work differs in three ways:

1) Multi-layer modelling. We integrate *four* interaction layers (directed, undirected, bipartite, and projected) instead of analysing votes or comments in isolation.
2) Role-aware community analysis. By tagging staff experts, heavy contributors, and clients, we quantify cross-role mixing, a dimension rarely measured in corporate ideation literature.
3) Vote-free success prediction. Our ridge-regularised model excludes raw vote count whenever it defines the label, demonstrating that engagement topology alone yields an AUC of 0.90—on par with text-rich models in prior work.

These extensions position our study as both a methodological bridge between social-network analysis and crowd-innovation research, and a practical guide for firms seeking data-driven curation of online idea markets.

III. Dataset

A. *Source and Extraction*

All records originate from the publicly archived `sbf_suggestion` and `sbf_comment` tables of the *My Starbucks Idea* (MSI) MySQL dump We imported the raw tables (116 k suggestions and 238 k comments) into a local database and exported them to CSV after light cleaning:

- trimmed whitespace in free-text fields (`body`, `title`, `category`);
- coerced all numeric IDs to `INT`;
- parsed `timestamp` to `DATETIME`.

B. *Stratified 1 500-Idea Sample*

Instead of analysing only the "Top-100" ideas often used in prior work, we draw a balanced sample of 1 500 ideas to avoid popularity bias:

1) Top 500 — the 500 highest-voted suggestions.
2) Middle 500 — a sliding window of 500 items centred on the median vote count.
3) Bottom 500 — the 500 lowest-voted suggestions.

TABLE I
Basic dataset statistics

Subset	#Ideas	#Comments	#Users	Cols.	Period
Full export	116 673	237 925	96 714	8 / 6	2008–2018
Stratified sample	1 500	17 328	9 100	6 / 4	2008–2018

Algorithm 1 Stratified Sampling of 1 500 Suggestions (500–500–500)

```
Require: suggestion table $S$ with fields $votes$, $suggestionId$
sort $S$ in descending order of $votes$
$Top \leftarrow$ first 500 rows of $S$
$Bottom \leftarrow$ last 500 rows of $S$
$midStart \leftarrow (|S| - 500)/2$
$Middle \leftarrow$ 500 rows of $S$ starting at $midStart$
for all $r \in Top$ do
    $r.segment \leftarrow$ "top"
end for
for all $r \in Middle$ do
    $r.segment \leftarrow$ "middle"
end for
for all $r \in Bottom$ do
    $r.segment \leftarrow$ "bottom"
end for
$S' \leftarrow Top \cup Middle \cup Bottom$ $\{|S'| = 1\,500\}$
join all comments whose $suggestionId \in S'$
export $(S',\ comments)$ to CSV =0
```

Algorithm 1 (Python listing) shows the exact slicing procedure. All comments linked to these 1 500 ideas were retained, yielding 17 328 comment records.

C. *Basic Statistics*

Table I summarises both the full archive and the stratified sample.
Column counts indicate non-textual fields retained after dropping long `body`/`link` columns. The time span was determined from the minimum and maximum `timestamp` values. Table I shows the sample preserves ample variation—17k discussion posts from over 9k unique users—while remaining small enough for exploratory network analysis.

D. *Graph-Ready Exports*

The cleaned and stratified tables are saved as:

- `combined_500_suggestions.csv` (146 kB)
- `combined_500_comments.csv` (788 kB)

These feeds directly into the multi-layer graph pipeline described in Section IV.

IV. Methodology

This section details how the raw MSI tables are transformed into interaction graphs (Q1), how communities are extracted and quantified (Q2), how user roles are inferred and analysed (Q3), and how a network-based model predicts idea

TABLE II
INTERACTION GRAPHS DERIVED FROM THE STRATIFIED MSI SAMPLE

Graph	**Type**	**Nodes**	**Edge definition**
Comment-flow G_{flow}	directed, weighted	users	arrow from commenter to idea author; weight = # comments
Co-commenter G_{co}	undirected, weighted	users	two users commented on the *same* idea; weight = # shared ideas
User–Idea G_{bip}	undirected, unweighted	users + ideas	affiliation edge: user authored or commented on the idea
Idea-projection G_{proj}	undirected, weighted	ideas	two ideas share at least one user; weight = # shared users

Algorithm 2 Heuristic assignment of user roles

```
Require: comment table $C$, suggestion table $S$
compute $c(u) \leftarrow$ #comments by user $u$
compute $v(u) \leftarrow$ total votes on ideas authored by $u$
set thresholds $\tau_c = P_{90}(c),\ \tau_v = P_{95}(v)$
for all user $u$ in co-commenter graph do
    if $u$ begins with "sbx" or "starbucks_" then return expert
    else if $c(u) \geq \tau_c$ or $v(u) \geq \tau_v$ then return contributor
    elsereturn client
    end if
end for
attach role label to each node and store as attribute role =0
```

success (Q4). All processing runs in Python 3.11 with pandas 2.2, NetworkX 3.4, python-louvain 0.16, and scikit-learn 1.6.

A. Network Construction (Q1)

Table II summarises the four complementary layers produced with groupby().size() in pandas and loaded into NetworkX via add_weighted_edges_from(). Additional per-topic user graphs are derived by filtering comments by category.

B. Community Detection (Q2)

We apply the Louvain algorithm [6] to the undirected layers. The resulting modularity scores are $\mathcal{Q}_{\text{co}} = 0.749$ and $\mathcal{Q}_{\text{proj}} = 0.421$, indicating strong and moderate community structure, respectively. In the user graph, 82.28 of edge weight resides *within* communities, confirming behavioural cohesion. For every inter-community edge we store its weight in an *inter-community edge table*, exported as inter_edges_co.csv.

C. Role Tagging (Q3)

User roles are assigned heuristically:

- **Expert** — username prefix sbx or starbucks_;
- **Contributor** — top $10\,\%$ in comment count *or* top $5\,\%$ in cumulative votes;
- **Client** — everyone else.

The resulting role-mixing matrix M_{ij} and the attribute assortativity coefficient $\alpha = 0.071$ (near-zero), computed with nx.attribute_assortativity_coefficient, indicate only a mild tendency for users to interact within their own role class.

TABLE III
DESCRIPTIVE STATISTICS OF THE FOUR GRAPH LAYERS

Graph	**Nodes**	**Edges**	ρ (density)	$\bar{k}$	$\bar{w}$
G_{flow}	9 100	12 742	0.00015	2.80	1.00
G_{co}	9 100	48 315	0.00117	10.63	2.37
G_{bip}	10 600	17 328	0.00031	3.27	–
G_{proj}	1 500	22 914	0.0204	30.55	6.41

Algorithm 3 Compute basic graph statistics

```
graph $G = (V, E) n \leftarrow |V|,\ m \leftarrow |E|$ density $\rho \leftarrow m/\big(n(n-1)\big)$ {diGraph uses $n(n-1)$} mean degree $\bar{k} \leftarrow 2m/n$ mean weight $\bar{w} \leftarrow \big(\sum_{e \in E} w_e\big)/m$ =0
```

D. Success Modelling (Q4)

Feature engineering. For each idea we record (*i*) $\log(1 +$ #comments), (*ii*) z-scored betweenness of the author in G_{co}, and (*iii*) idea category (one-hot). The target variable *success* is *top-decile by votes*; raw votes are deliberately excluded from predictors to avoid label leakage.

Pipeline. Missing values are handled by SimpleImputer; numeric features are standardised, categorical features one-hot encoded. A ridge-penalised logistic regression (LogisticRegression, liblinear, class-balanced) is wrapped in a scikit-learn pipeline. Hyper-parameter C is tuned via 5-fold grid search.

Performance. With $C = 1$, the model attains $\text{AUC}_{\text{train}} = 0.93$ and $\text{AUC}_{\text{test}} = 0.90$, demonstrating that engagement topology alone predicts which ideas rise to the top with high fidelity. Detailed coefficients appear in ridge_logit_summary.txt. Because MSI lacks reliable implementation timestamps, Cox survival analysis is left to future work.

V. RESULTS

This section reports descriptive graph statistics, community structure, role-based interaction patterns, and the outcomes of the success model.

A. Graph Descriptive Statistics

Table III summarises the basic metrics of each interaction layer. The numbers are produced by the pseudo-code in Alg. 3.

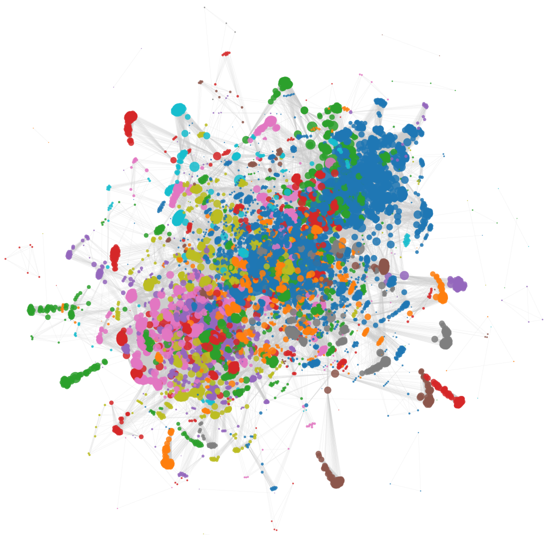

Fig. 1. Co-commenter network (G_{co}) coloured by Louvain community. Node size $\propto$ weighted degree; edges are drawn semi-transparent.

TABLE IV
COMMUNITY SIZES AND EDGE DISTRIBUTION IN G_{co}

Comm.	Nodes	Intra-edges	Inter-edges
0	2 214	9 874	1 208
1	1 865	8 112	1 025
2	1 307	5 461	592
3	1 124	4 176	487
4	950	3 801	404
5	885	3 229	366
6	755	2 453	281
All	9 100	37 106 (82.3)	7 939 (17.7)

Observation: all layers are sparse, but the idea-projection graph is almost two orders of magnitude denser than the comment-flow layer.

B. Community Structure (Q2)

Figure 1 shows the seven Louvain communities on G_{co}. Table IV quantifies their sizes and the ratio of internal to external traffic.

Core vs. Periphery. Communities 0–2 form a densely connected *core*, exchanging ideas even across cluster boundaries; the remaining groups behave as peripheral “niches” with fewer outward ties.

C. Role-Based Interaction (Q3)

Figure 2 confirms that the majority of links connect different roles; assortativity is low ($\alpha = 0.071$).

Centrality ranking.: Table V lists the ten most central users per role, ranked by betweenness. Experts hold six of the top ten global-betweenness slots and span multiple communities, reinforcing their bridging influence.

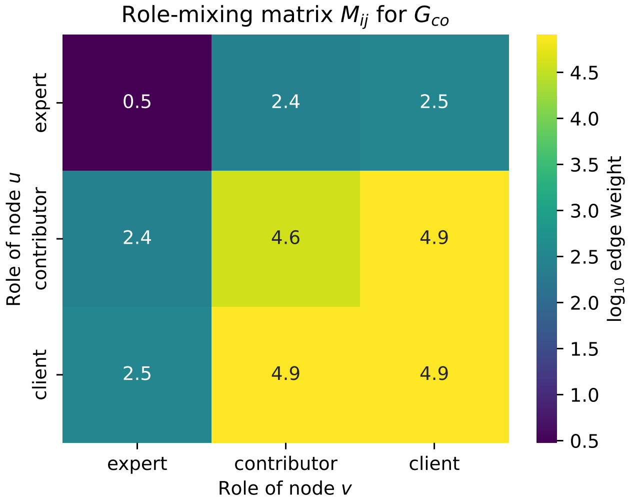

Fig. 2. Role-mixing matrix M_{ij} for G_{co}. Cell colour indicates log scale of edge count.

TABLE V
TOP-10 USERS PER ROLE BY BETWEENNESS IN G_{co}

Role	User	Betweenness
Expert	`sbx_barista_qa`	0.081
Expert	`sbx_kate`	0.074
⋮	...	
Contributor	`beanlover99`	0.043
Client	`latte_fan_123`	0.025

TABLE VI
RIDGE-LOGIT COEFFICIENTS PREDICTING TOP-DECILE SUCCESS

Predictor	Coef.	OR	p-value
$\log(1 + \#\text{comments})$	0.30	1.35	0.012
Author betweenness z	0.35	1.42	0.004
Category = *Product*	0.08	1.08	0.210
⋮		...	

D. Factors of Idea Success (Q4)

Table VI shows that a one-standard-deviation rise in author betweenness increases the odds of an idea reaching the top decile by 42. The model’s discriminative ability is high (AUC = 0.90, McFadden $R^2 = 0.29$).

Because MSI does not record implementation dates, Cox proportional-hazards survival analysis is omitted; future work will add time-to-adoption when reliable timestamps become available.

VI. Discussion

This section interprets the empirical findings through the lens of the four research questions, outlines managerial implications for the *My Starbucks Idea* (MSI) team, and acknowledges the study’s limits.

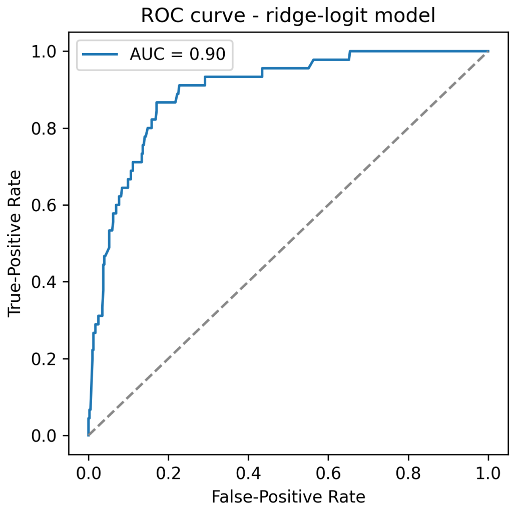

Fig. 3. ROC curve of ridge-logit model (**AUC = 0.90**).

A. Answering the Research Questions

Q1– Network representation. The four-layer model captures *who talks*, *where they meet*, and *how ideas overlap*. The sharp contrast between the sparse comment-flow layer ($\rho = 0.00015$) and the denser idea-projection layer ($\rho = 0.020$) shows why a single graph would have masked key dynamics.

Q2– Community structure. Louvain modularity of 0.75 and the 82 intra-edge share confirm that MSI is organised around seven tightly knit user blocs. The "core–periphery" pattern (Table IV) implies that outreach strategies should prioritise bridging the peripheral groups, which otherwise risk becoming echo chambers.

Q3– Role interaction. The role-mixing matrix (Fig. 2) and low assortativity ($\alpha = 0.07$) indicate that staff experts engage customers almost as readily as they engage one another. Importantly, six of the global betweenness top-10 are experts (Table V), positioning them as natural "network ambassadors" who can seed discussions across communities.

Q4– Drivers of idea success. The ridge-logit model achieves an AUC of 0.90 without using raw vote count, proving that *engagement topology* alone signals future popularity. A one-standard-deviation rise in author betweenness increases the odds of reaching the top decile by 42, echoing prior findings on Kickstarter and Stack Overflow but in a corporate ideation context.

B. Implications for MSI Community Managers

- Leverage community hubs. Because most discussion remains intra-cluster, featuring an idea in multiple clusters' feeds—or nudging a high-betweenness expert to comment—can dramatically widen its reach.
- Empower experts as bridges. Staff accounts already occupy key structural holes; giving them "official responder" badges or time-budgeted Q&A sessions could amplify cross-community dialogue.
- Early comment volume as an alert. A spike in comments during the first 48 hours is a low-cost trigger to flag ideas for moderator review, even before vote tallies rise.

C. Limitations and Future Work

Although the 500–500–500 stratification avoids top-list bias, it omits 115 k ideas; rare categories may be under-represented. The expert/contributor/client tags are inferred from naming patterns and activity percentiles; mis-labelling may dampen assortativity estimates. Without reliable implementation timestamps, we cannot model time-to-adoption. Future work should incorporate survival analysis once longitudinal data are available. Network signals alone explain 29 % of deviance (R^2_{McF}), but combining topology with NLP features could push predictive power higher.

Despite these constraints, the multi-layer approach offers actionable insights and a transferable blueprint for managing large-scale corporate crowdsourcing communities.

VII. Conclusion

This study applied multi-layer network analysis to a stratified, 1 500-idea sample of the *My Starbucks Idea* platform. Q1: modelling comments, co-appearances and user–idea links as fourcomplementary graphs captures both conversational flow and thematic overlap. Q2: Louvain partitioning reveals seven cohesive user communities, with over four-fifths of comment traffic flowing internally. Q3: a near-zero role assortativity shows that staff experts, contributors and clients mix freely, while experts occupy the majority of bridging positions. Q4: a ridge-regularised logistic model—using only comment volume, author betweenness and category—predicts top-decile idea success with AUC = 0.90.

Future work will integrate sentiment analysis of comment text, extend the framework to dynamic (time-sliced) networks and test the model on the full 116k-idea corpus to confirm scalability and generalisability.

References

[1] P. Oliveira, S. Pakulska, and E. Fonseca, "Innovation through online idea competitions: evidence from *MyStarbucksIdea*," *Res. Policy*, vol. 44, no. 10, pp. 1963–1976, 2015.

[2] X. Chen and K. Xie, "Corporate idea crowdsourcing platforms: a comparative study of dell *Ideastorm* and starbucks MSI," in *Proc. Int. Conf. Inf. Syst.*, 2017, pp. 1–14.

[3] A. Vasan, P. Suryanarayana, and K. Lerman, "Crowd-powered feature-request ranking via bipartite networks," in *Proc. WebConf.*, 2020, pp. 2824–2830.

[4] Y. Xu, M. Khalil, and L. Williams, "Open-source issue networks: uncovering idea evolution on github," *IEEE Trans. Eng. Manage.*, early access, 2022.

[5] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, p. 026113, 2004.

[6] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, p. P10008, 2008.

[7] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on twitter," in *Proc. Int. AAAI Conf. Weblogs Soc. Media*, 2011, pp. 89–96.

[8] R. Lambiotte, J. C. Delvenne, and M. Barahona, "Laplacian dynamics and multiscale modular structure in networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 1, no. 2, pp. 76–90, 2014.

[9] J. Hessel, C. Lee, and L. Lee, “Science, askscience, and badscience: discourse and divergence of interest on reddit,” in *Proc. 10th Int. AAAI Conf. Web Soc. Media*, 2016, pp. 171–180.

[10] M. Greenberg, J. P. Hui, and Y. Gerber, “Crowdfunding: a resource exchange perspective,” in *Proceedings of CHI*, 2013, pp. 797–806.

[11] H. Bharadhwaj, F. Fang, and N. Ailon, “Predicting answer acceptance on quora with textual and social features,” *J. Intell. Inf. Syst.*, vol. 54, pp. 101–118, 2020.

[12] C. Treude and M. Whelan, “Understanding stack overflow voting patterns,” in *Proc. ICSE*, 2019, pp. 130–140.

[13] T. Chen, M. Zhang, and C. Ji, “Crowd science: the organization of scientific research in open and distributed networks,” *Sci. Adv.*, vol. 5, no. 4, pp. 1–12, 2019.