Network Optimization Models for Relationship Building in Venture Capital

Vineet Reddy, Avya Kalra, Era Ahsan

Abstract

Sequoia Capital has publicly discussed their proprietary data science system that functions as an intelligence engine for the startup ecosystem. A core component is a "Talent Map" that acts as a ranking algorithm for people, built by systematically asking high-performing operators to identify the most talented peers in their networks. The result is a web of endorsements connecting thousands of individuals across the technology industry.

This talent map serves two strategic purposes. First, it allows the firm to assess startup team quality with high precision: investors can query the system to determine the percentile ranking of a company's engineering team. Second, the system supports tracking talent before companies are formed, through the seed stage, and for late-stage investment.

As these talent maps grow, two optimization challenges emerge. First, the conventional networking strategy of reaching out to prominent individuals leads to significant redundancy: the same clusters are discovered repeatedly. Second, relationships require active maintenance to remain valuable. This project applies IEOR 240 optimization techniques to address both challenges, implementing an exact integer programming solution using Gurobi and a greedy approximation algorithm.

1. The "Sentinel" Problem (Avoiding Redundancy)

The standard approach to network expansion is connecting with high-profile individuals and reaching out to people they recommend. However, this produces highly redundant networks: prominent contacts tend to know overlapping sets of people.

We hypothesize that there exist "underappreciated" nodes that provide access to high-density talent pools, and "Sentinels" who are well-connected to these nodes and can unlock otherwise unreachable talent. Our goal is to identify the optimal set of Sentinels such that total unique talent discovered is maximized.

IEOR 240 Concept

We model this as a Maximum Coverage Problem. The objective is to select k Sentinels such that the union of talent they collectively know is maximized. If Sentinel A and B both know the same 50 engineers, selecting both provides no additional coverage beyond selecting one.

Our Formulation (Integer Program)

Sets:

- I: Set of potential people to reach out to (Sentinels)
- J: Set of "Talent" (the valuable connections to discover)

Parameters:

- $a_{ij} = 1$ if Sentinel i knows Talent j, 0 otherwise
- w_j: Quality score for Talent j (mean trust rating received)
- K: Budget constraint (maximum number of Sentinels to contact)

Decision Variables:

- $x_i \in \{0, 1\}$: 1 if Sentinel i is selected for outreach
- $y_j \in \{0, 1\}$: 1 if Talent j is covered (discovered) through selected Sentinels

Objective:

$$\text{Maximize} \sum_{j \in J} w_j y_j$$

Constraints:

1. Talent j can only be covered if at least one selected Sentinel knows them:

$$y_j \leq \sum_{i \in I} a_{ij} x_i \quad \forall j \in J$$

2. At most K Sentinels may be selected:

$$\sum_{i \in I} x_i \leq K$$

3. Binary constraints: $x_i, y_j \in \{0, 1\}$

This BIP formulation uses Branch and Bound/Cut and explicitly solves the redundancy problem because the objective counts y_j once regardless of how many Sentinels know that person.

2. The "Maintenance" Problem (Relationship Decay)

Professional relationships require active maintenance. A contact valuable two years ago may no longer be accessible if relationship has gone dormant. The optimization challenge is determining which relationships to prioritize for re-engagement, given limited time.

IEOR 240 Concept

We model this as a Knapsack Problem where the value of reconnecting with each contact depends on staleness, talent quality, and network centrality (degree).

Our Formulation
Parameters:

- t_i: Days since last contact with person i
- w_i: Talent score (mean trust rating received) for person i
- d_i: Degree of person i (number of connections)
- c_i: Cost (time in minutes) to reconnect with person i
- T: Total time budget available

Value Function:

$$v_i = \log(t_i + 1) \cdot w_i \cdot \sqrt{d_i}$$

The logarithm provides diminishing urgency with staleness; the square root on degree prevents over-weighting high-degree nodes.

Optimization Model:

$$\text{Maximize} \quad \sum_i v_i \, x_i$$
$$\text{Subject to} \quad \sum_i c_i x_i \leq T$$
$$x_i \in \{0, 1\}$$

3. New Method: Submodular Optimization (Greedy Algorithm)

To satisfy the requirement of using a method not covered in class, we implement a greedy algorithm based on submodular optimization theory, comparing it against the exact Gurobi IP solution.

Motivation

The Maximum Coverage problem is NP-Hard. While Gurobi solves it optimally, this may become prohibitive for large networks. We implement an efficient approximation with provable guarantees.

The Greedy Approach

At each iteration, select the Sentinel who adds the most new, unique weight to the covered talent pool. Repeat until K Sentinels are selected.

Theoretical Guarantee

The coverage function is submodular, meaning it exhibits diminishing returns. The theorem by Nemhauser, Wolsey, and Fisher (1978) proves the greedy algorithm guarantees a solution within $(1-1/e)$, or about 63%, of optimal for submodular maximization subject to cardinality constraints.

Comparison Metrics

We compare the greedy algorithm against the optimal Gurobi IP solution on:

- Total talent weight covered by selected Sentinels
- Runtime comparison across varying problem sizes
- Empirical ratio of greedy solution to optimal solution

4. Data and Computational Approach

We use a real-world social network dataset from the Stanford Large Network Dataset Collection (SNAP) to validate our optimization models.

Dataset: Bitcoin Alpha Trust Network

We use the Bitcoin Alpha weighted signed directed network[1], capturing trust relationships between Bitcoin Alpha OTC platform users. Members rate each other on a scale of -10 (total distrust) to $+10$ (total trust). The dataset contains 3,783 nodes and 24,186 edges, with 93% positive ratings. Each edge includes a trust rating and Unix timestamp.

Data Adaptation: Mapping to the VC Talent Map

The Bitcoin Alpha trust network provides a compelling analog to a VC talent map. We interpret the data as follows:

- Each trader (node) represents a professional contact, and rating another trader mirrors vouching for a peer.
- The signed ratings $[-10, +10]$ represent endorsement strength.
- Talent scores (w_j) are computed as the mean trust rating received, capturing aggregate reputation.
- We convert the graph to undirected since professional relationships are mutual.
- A relationship between i and j means Sentinel i knows Talent j (coverage relationship a_{ij}).
- Real timestamps enable meaningful analysis of relationship decay.

---

[1]S. Kumar et al., IEEE ICDM 2016; ACM WSDM 2018. Available at snap.stanford.edu/data/soc-sign-bitcoin-alpha.html.

Synthetic Contact Enrichment

The Bitcoin Alpha dataset contains only anonymous node IDs with no personal information. To create a realistic demo that resembles an actual VC talent map with contact details, we enriched the graph with synthetic person data from `people-10000.csv`[2]. Each node ID in the graph is mapped to the corresponding index in the people CSV, attaching fields including name, email, phone, and job title. This enrichment is purely for demonstration purposes and does not affect the optimization results, which depend only on graph structure and trust ratings.

Visualization Subgraph

The complete graph (3,783 nodes, 14,124 edges) is too large for browser rendering. We select a subgraph of 500-600 nodes: all K Sentinel nodes, all Maintenance-selected nodes, top 100 talent nodes, and up to 10 neighbors of each important node.

5. Results

We ran our optimization pipeline on the full Bitcoin Alpha network (3,783 nodes, 14,124 edges) with $K = 10$ Sentinels and a maintenance budget of $T = 2800$ minutes.

Sentinel Problem Results

Method	Coverage	Runtime (s)	% of Optimal
IP (Optimal)	2417.56	0.1158	100.0%
Greedy	2181.27	0.0289	90.2%
Naive (Top Degree)	1569.46	0.0024	64.9%

The IP formulation achieves 54.0% higher coverage than the naive baseline, confirming our hypothesis that simply targeting high-degree nodes leads to significant redundancy. The greedy algorithm achieves 90.2% of optimal coverage while running 4x faster than the IP solver, well above the theoretical 63% guarantee. Both optimization approaches substantially outperform the naive strategy.

Maintenance Problem Results

Metric	Value
Time Budget	2800 minutes
Relationships Selected	97
Total Maintenance Value	9710.13
Average Dormancy	924 days

The optimization selected 97 high-priority relationships for re-engagement. The average dormancy of 924 days (over 2.5 years) shows the model prioritizes relationships that have been neglected longest while still providing high network value.

---

[2]Available at `datablist.com/learn/csv/download-sample-csv-files`.