# Link analysis on customer shopping networks

James Wang

December 15, 2016

Technical Case Interview

# Background

*The client, a leading global specialty food retailer, has developed a novel approach to identifying customers who shop together.*

*They would like to improve the efficiency of promotional spend via the loyalty program, and have provided us network data of their Seattle customers to explore.*

# Roadmap

- High-Level Methodology
- Link Analysis Results I: Revenue Analysis
- Link Analysis Results II: Relationships Analysis
- Business Recommendations
- Concluding Thoughts

# High-Level Methodology

# Defining the OKRs

- Primary Objective
  - Our primary point of contact is the CMO, whose goal is to improve the efficiency of promotional spend via the existing loyalty program.

# Defining the OKRs

- Primary Objective
  - Our primary point of contact is the CMO, whose goal is to improve the efficiency of promotional spend via the existing loyalty program.
- Key Results Needed
  1. *Define* efficiency of spend
  2. *Quantify* efficiency of spend
  3. *Implement* efficient spending strategy

# Defining the OKRs

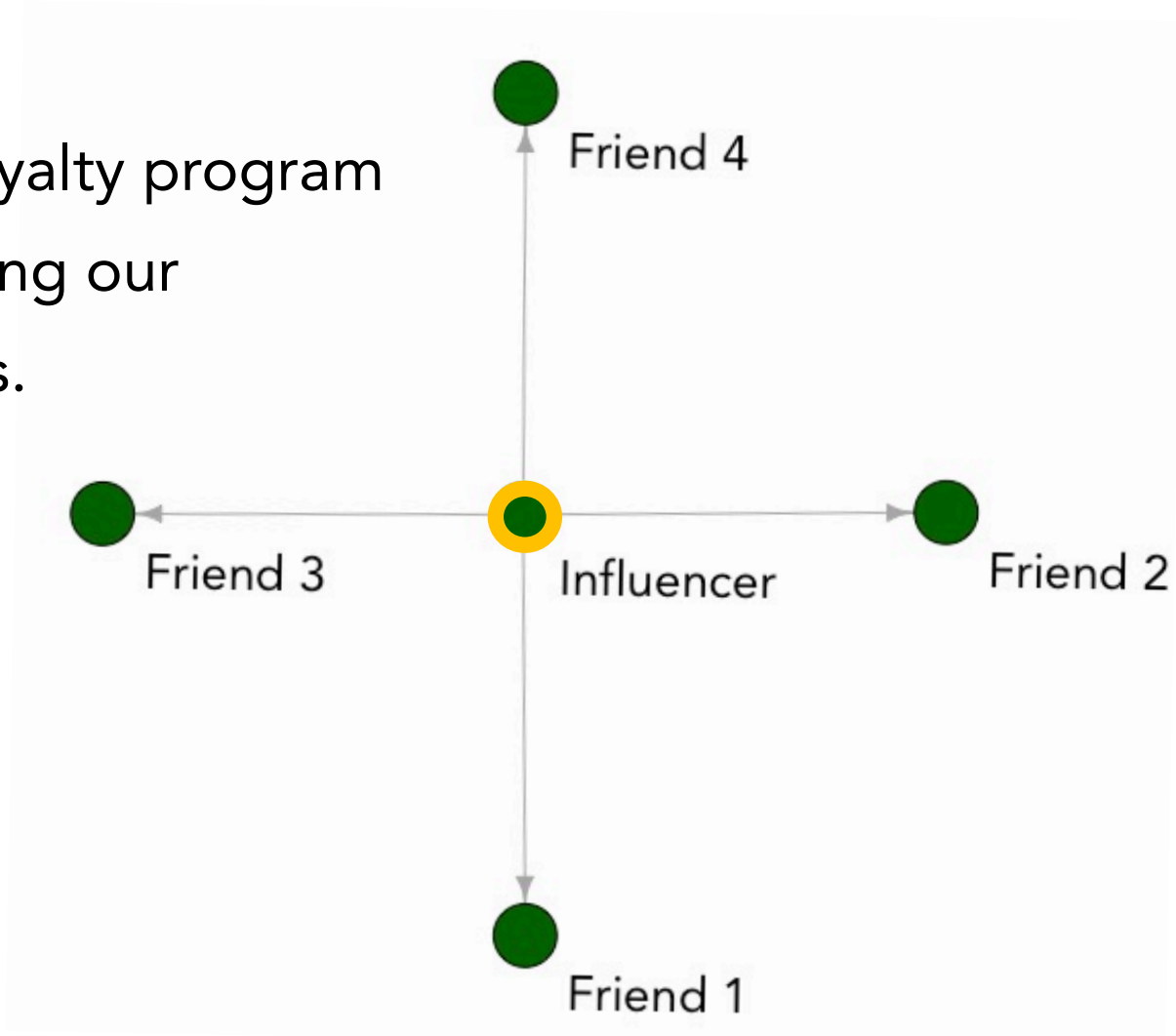
- Secondary Objectives
  - Explore data on full-graph level and zipcode-level
  - Deep dive into customer relationships

# Defining the OKRs

- Secondary Objectives
  - Explore data on full-graph level and zipcode-level
  - Deep dive into customer relationships
- Key Results Needed
  - Are there links between the size / strength of networks and certain zip codes and neighborhood-types in Seattle?
  - How balanced are relationships?
    - Are customers often paired with others who have similar number of friends, or is it typically one sided?
    - Similarly for revenue – big spenders with big spenders or not?

# 1. Define efficiency of spend

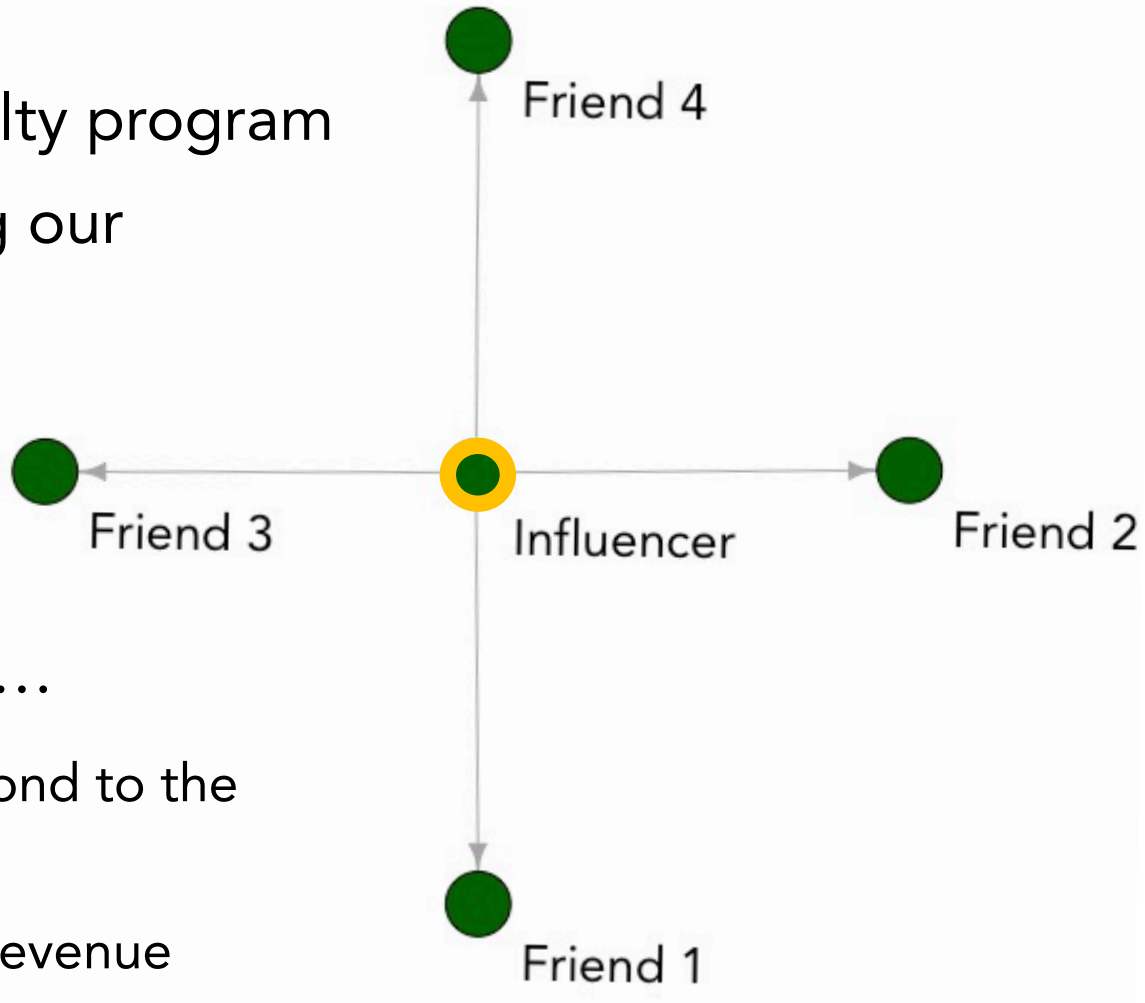
Maximize ROI on “loyalty program spending” by targeting our high-value customers.




BCG

# 1. Define efficiency of spend

Maximize ROI on "loyalty program spending" by targeting our high-value customers.




A **high-value** customer…

1. …is most likely to respond to the loyalty program
2. …generates the most revenue

BCG

# 2. Quantify efficiency of spend

Who **should** we be targeting?

Identify our high-value customers using network learning techniques

Who are our high-value customers?

# 2. Quantify efficiency of spend

| Who **should** we be targeting? | Who **are** we targeting? |
| --- | --- |
| Identify our high-value customers using network learning techniques | Pull data on customers who have been sent loyalty program promotions |
| Who are our high-value customers? | Are they our high-value customers? |

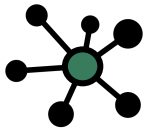
# 2. Quantify efficiency of spend

| Who **should** we be targeting? | Who **are** we targeting? |
| --- | --- |
| Identify our high-value customers using network learning techniques | Pull data on customers who have been sent loyalty program promotions |
| Who are our high-value customers? | Are they our high-value customers? |

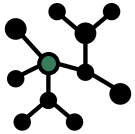
Are we targeting the right customers with our loyalty program?

# Identifying high-value customers

*Tactic 1: Customers who **individually** spend the most*



*Tactic 2: Customers whose **1st degree connections** spend the most*



*Tactic 3: Customers whose **sub-graph** spends the most*

*Tactic 4: **Zip codes** who spend the most*

# 3. Implement efficient spending strategy

## Test effectiveness of new strategy with experiment

Define metric to track and forecast metric performance

Monthly Revenue ($)

Identify high-value customers

Set up treatment and control groups

Run experiment for 3-6 months

Evaluate results

High-value customers

Treatment

Control

Low-value customers

Treatment

Control

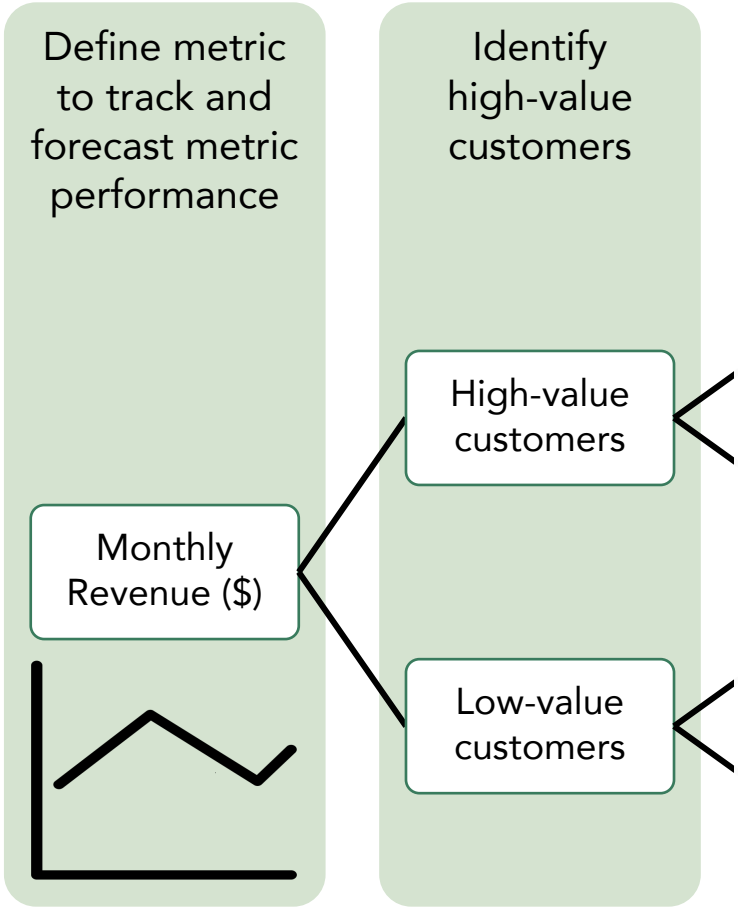
Track metric

Causal Inference:

Is difference in actual metric performance and forecasts statistically significant?

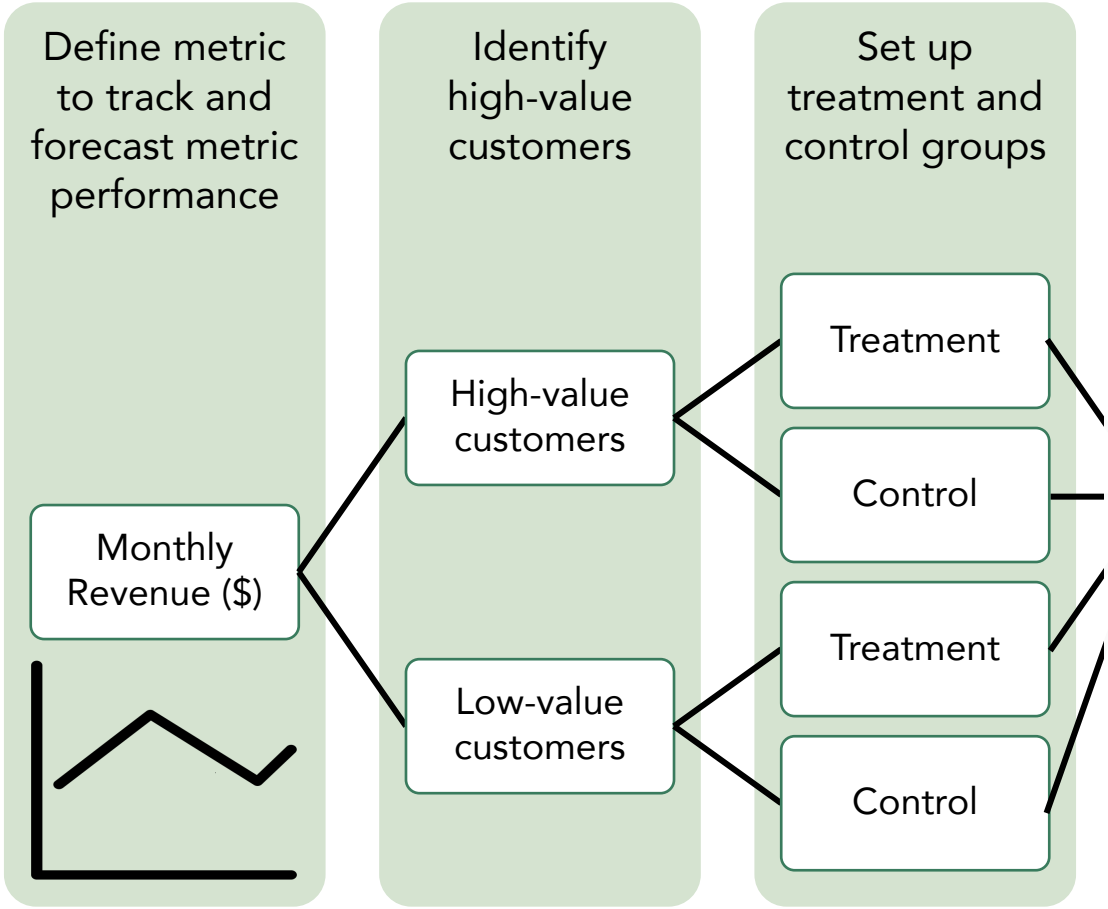
# 3. Implement efficient spending strategy

## Test effectiveness of new strategy with experiment



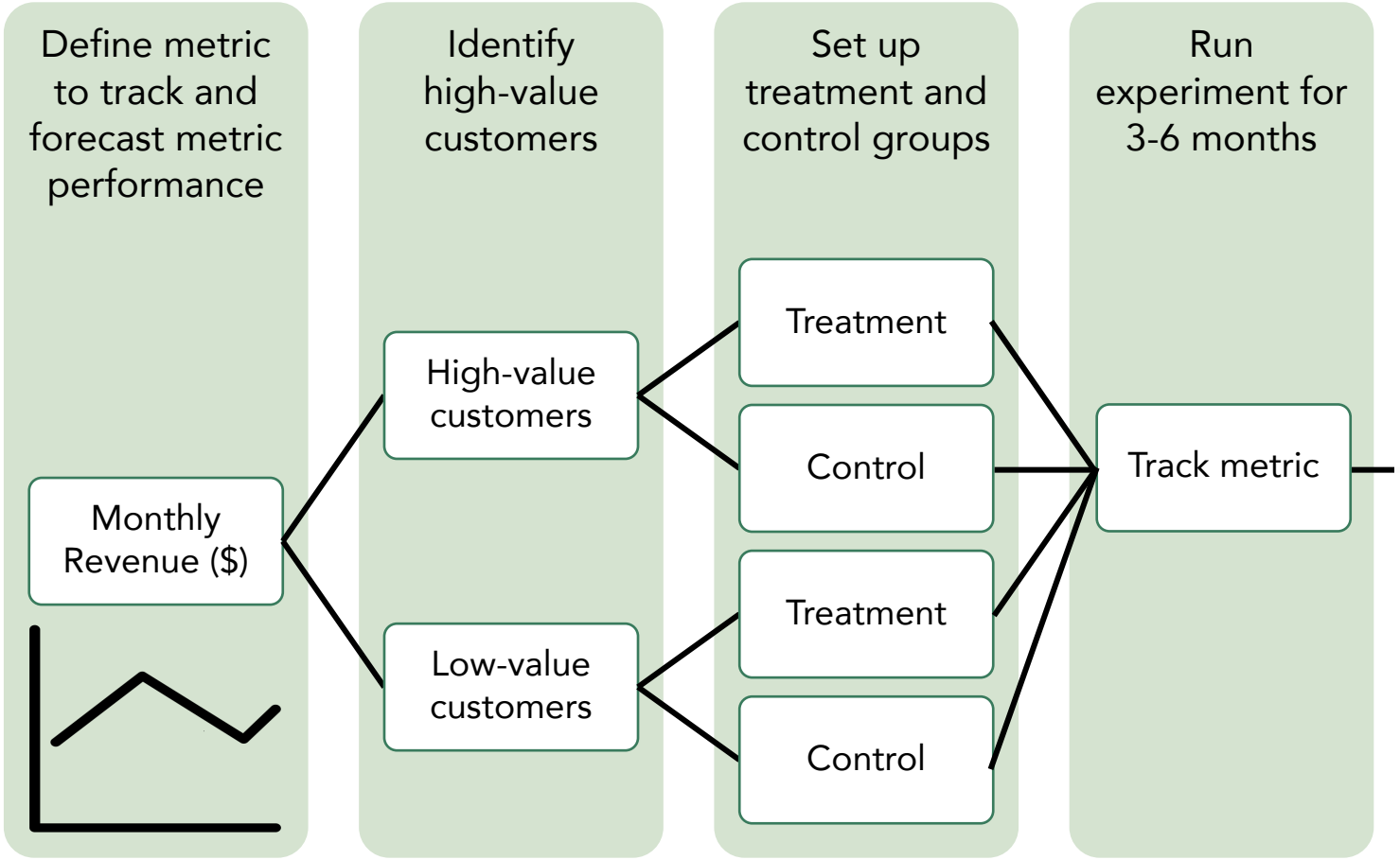
# 3. Implement efficient spending strategy

## Test effectiveness of new strategy with experiment



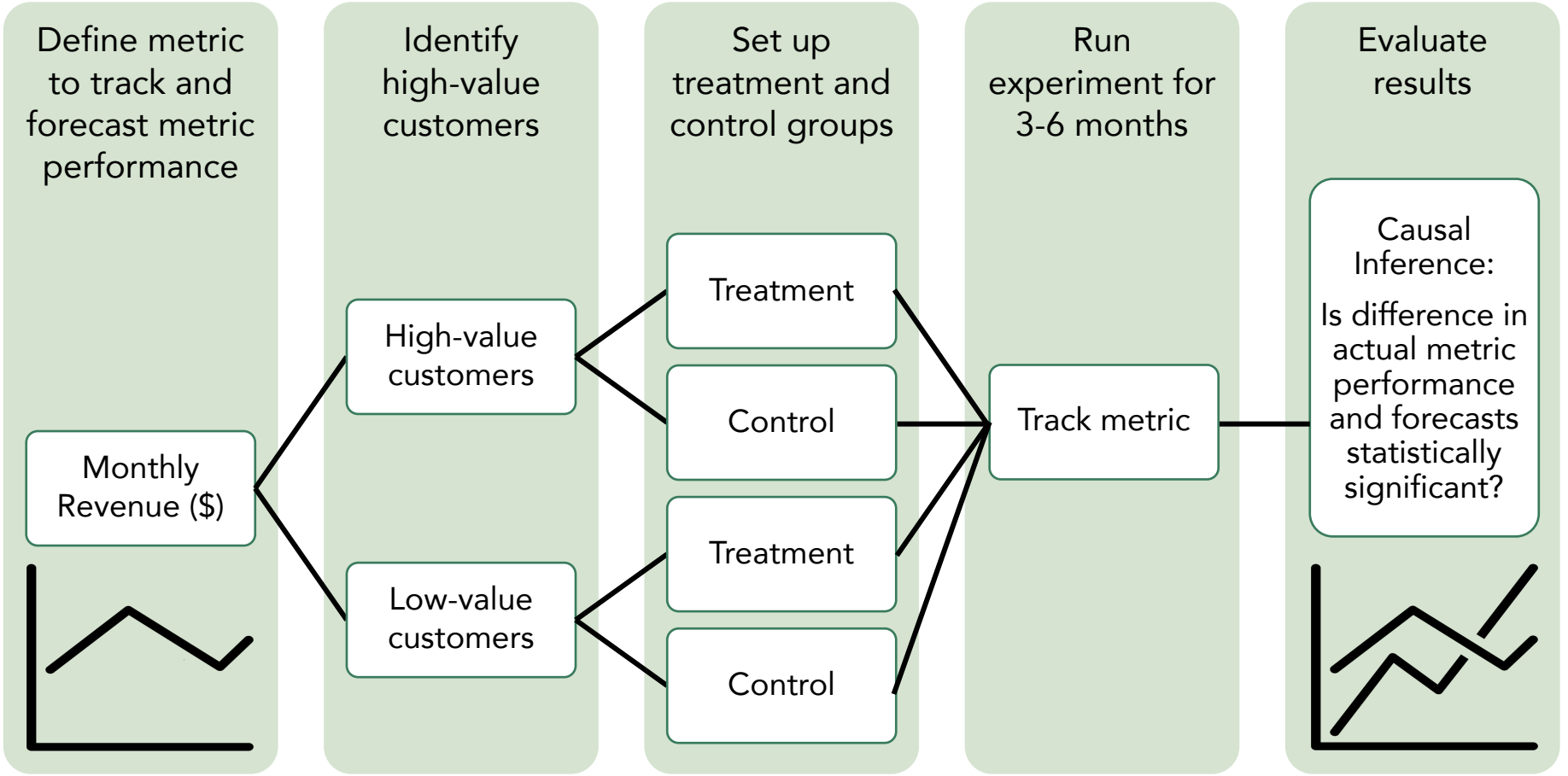
# 3. Implement efficient spending strategy

## Test effectiveness of new strategy with experiment



# 3. Implement efficient spending strategy

## Test effectiveness of new strategy with experiment



# Link Analysis Results I

Identifying high-value customers in terms of revenue

# Properties of the data

- Raw Data
  - List of customer shopping duos (n = 7792) and corresponding strength of relationship (1 to 10) and zip codes (n = 26)
  - List of customer IDs and their spending (n = 9628)

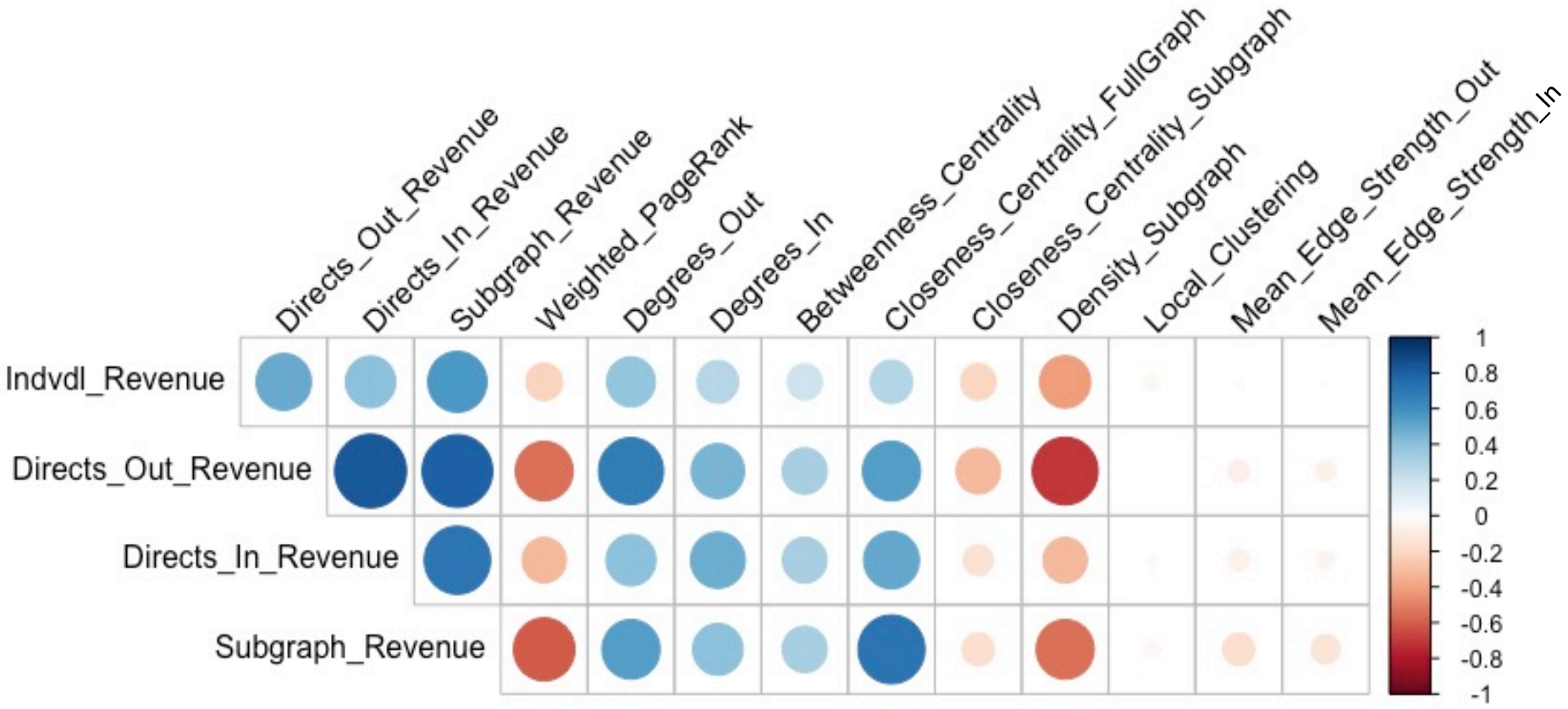
# Properties of the data

- Raw Data
  - List of customer shopping duos (n = 7792) and corresponding strength of relationship (1 to 10) and zip codes (n = 26)
  - List of customer IDs and their spending (n = 9628)
- Features to create
  - Measures of revenue, measures of node centrality, measures of connectedness, zip code demographics

# Properties of the data

- Raw Data
  - List of customer shopping duos (n = 7792) and corresponding strength of relationship (1 to 10) and zip codes (n = 26)
  - List of customer IDs and their spending (n = 9628)
- Features to create
  - Measures of revenue, measures of node centrality, measures of connectedness, zip code demographics
- Transformations of features
  - Log certain features (e.g. revenue) to normalize

# Correlation Analysis




n = 9628

# Tactic 1: Individual Revenue

| Most important factors* | |
|---|---|
| Name | Relationship |
| Subgraph Revenue | +++ |
| Closeness (Full Graph) | – – |
| Weighted PageRank | ++ |
| Density (Subgraph) | – – |

| High-Value Customers | |
|---|---|
| 3153 | 2336 |
| 3354 | 2357 |
| 1534 | 2931 |
| 1549 | 19 |

Insight 1:
*The customers who spend the most are typically part of high spending, low density subgraphs. Interestingly, the more "close" a customer is to all other customers in the full graph, the less they spend.*

Insight 2:
*Customers spend more when they are connected with more friends that pull them to shop (and in turn, those friends are pulled by their own friends to shop, etc). These customers are "popular" in their circles.*

*Cross-validated (5-fold) L1-regularized OLS regression and random forest regression, n = 1880, p = 13

# Tactic 2a: Directs (Out-Degree) Revenue

| Most important factors* | |
|---|---|
| Name | Relationship |
| Directs Revenue (In-Degree) | ++ |
| Subgraph Revenue | ++ |
| Density (Subgraph) | – – |
| Closeness (Full Graph) | – – |

| High-Value Customers | |
|---|---|
| 1816 | 2336 |
| 2926 | 1971 |
| 2931 | 3560 |
| 2928 | **1814** |

Insight 1:
*Combined revenue of a customer's friends (out) is positively related with the combined revenue of a customer's friends (in), even though the customer him/herself may not be a high-spender. This relates to subgraph revenue.*

Insight 2:
*Again, it seems that if your subgraph is not as highly connected (everyone is friends with everyone else), there is more spending. Furthermore, it seems customers that are closest to everyone on the graph also have friends who have lower spending.*

*Cross-validated (5-fold) L1-regularized OLS regression and random forest regression, n = 1880, p = 13

# Tactic 2b: Directs (In Degree) Revenue

| Most important factors* | |
|---|---|
| Name | Relationship |
| Directs Revenue (Out-Degree) | +++ |
| Subgraph Revenue | ++ |
| Density (Subgraph) | ++ |
| Closeness (Full Graph) | + |

| High-Value Customers | |
|---|---|
| 1434 | 249 |
| 3634 | 252 |
| 1859 | 3552 |
| 3388 | 4191 |

Insight 1:
*The friends who bring a customer to shop are high-spending if the customer shops with other high-spenders, and additionally, is part of a high-spending subgraph.*

Insight 2:
*Surprisingly, unlike the previous two tactics, being part of a customer's higher density subgraph means more spending. Furthermore, higher closeness in regards to the whole graph means the more people the customer knows, the more those who want to shop with the customer will spend. This reinforces the popularity notion.*

*Cross-validated (5-fold) L1-regularized OLS regression and random forest regression, n = 1880, p = 13

# Tactic 3: Subgraph Revenue

## Most important factors*

| Name | Relationship |
|---|---|
| Closeness (Full Graph) | +++ |
| Directs Revenue (Out-Degree) | ++ |
| Weighted PageRank | – – |
| Directs Revenue (In-Degree) | + |
| Individual Revenue | + |

## High-Value Customers

| | |
|---|---|
| 1813 | 1840 |
| 1823 | 1972 |
| 1816 | 3361 |
| **1814** | 929 |

Insight 1:
*As expected, the other revenue types all positively relate to subgraph revenue quite strongly.*

Insight 2:
*The closer the customer is to the rest of the graph, the more their subgraph spent.*

*Cross-validated (5-fold) L1-regularized OLS regression and random forest regression, n = 1880, p = 13

# Tactic 4: Average Zipcode Revenue

## Most important factors*

| Name | Relationship |
|---|---|
| Population Density | – |
| Median Age | + |

Insight 1:
*Network attributes (density of zipcode, average relationship strength) did not appear to be important.*

## High-Value Zipcodes

*See customer targeting database.*

Insight 2:
*Demographic attributes are weakly related with average customer spend by zipcode. More spending occurred in less populated zip codes and areas with a higher median age.*

*Cross-validated (5-fold) L1-regularized OLS regression and random forest regression, n = 23, p = 8

# What can we conclude from insights?

- The different types of revenue are very important when explaining one another, especially **subgraph revenue**.
- Some combination of node **closeness** in the full graph, weighted **PageRank**, and **density** of a node's entire subgraph are key drivers of revenue.
- With n = 23 and weak importance, more data is needed to conduct an intra-zipcode level analysis.
- See Appendix for inter-zipcode level analysis.

# What can we conclude on customers?

- Customers whose subgraphs spend highly are more likely to spend highly themselves. This type of effect may be attributed to socioeconomic reasons.
- Density, closeness, and weighted PageRank seemed to flip-flop (in terms of direction of relationship) between different types of revenue. We may want to defer to domain expertise or customer surveys to better understand this behavior.

# What can we conclude on customers?

| node | Individual Revenue | Out Revenue | In Revenue | Subgraph Revenue |
|---|---|---|---|---|
| 1813 | 1127.96103 | 15583.62864 | 4864.8919 | 308916.6078 |
| 1823 | 903.5284028 | 8473.56533 | 4149.7573 | 221728.9869 |
| 1816 | 1275.713911 | 21858.00195 | 4955.623 | 220071.504 |
| 1814 | 1007.821746 | 17329.8239 | 4002.7294 | 198798.0931 |
| 1840 | 571.0716826 | 5077.359676 | 5309.7363 | 191168.8063 |
| 1972 | 1328.154131 | 15808.71895 | 4779.817 | 184960.3214 |
| 3361 | 681.017972 | 3618.079414 | 2411.2574 | 158741.3565 |
| 929 | 929.0759047 | 9132.060587 | 3365.9582 | 149932.0619 |
| 252 | 481.7915537 | 2819.750242 | 5884.7909 | 140459.6311 |
| 3393 | 533.1652893 | 2300.454359 | 4544.3342 | 139159.8834 |

# Link Analysis Results II

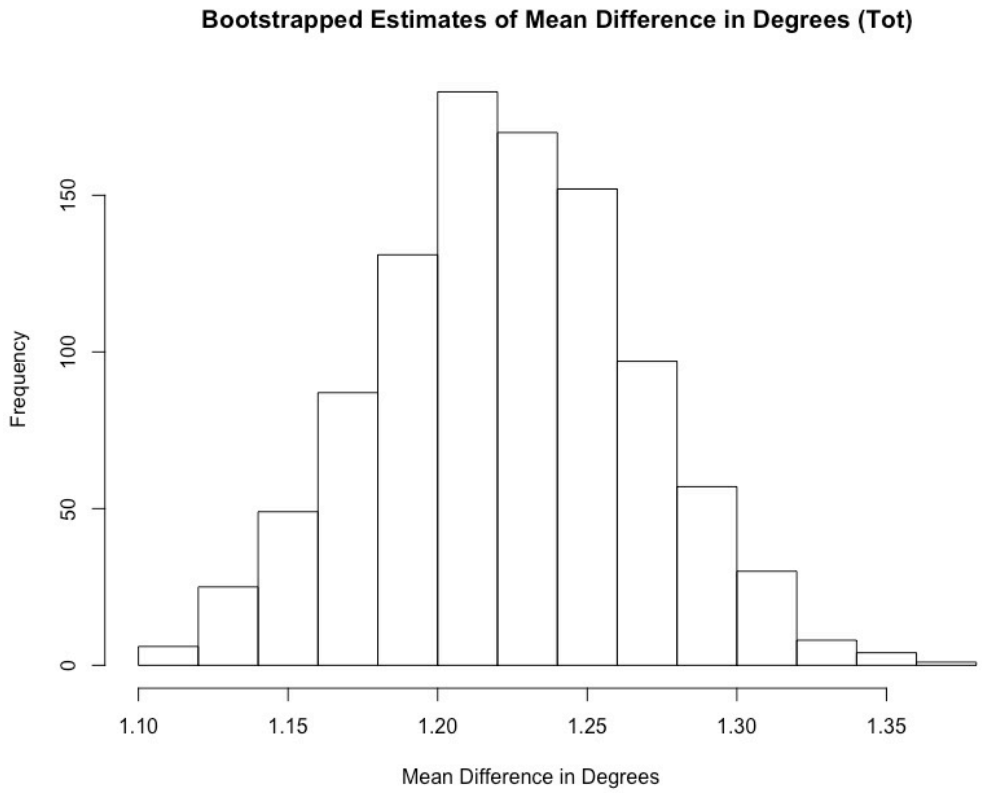
Investigating relationships between our customers

Are customers often paired with others who have **similar number** of friends, or is it typically one sided?

## **Total** Degrees

- Result: Customers typically have 1-2 more total friends *than* the friends they interact with.

Insight:
*Customers typically have the same network size (amount of friends).*

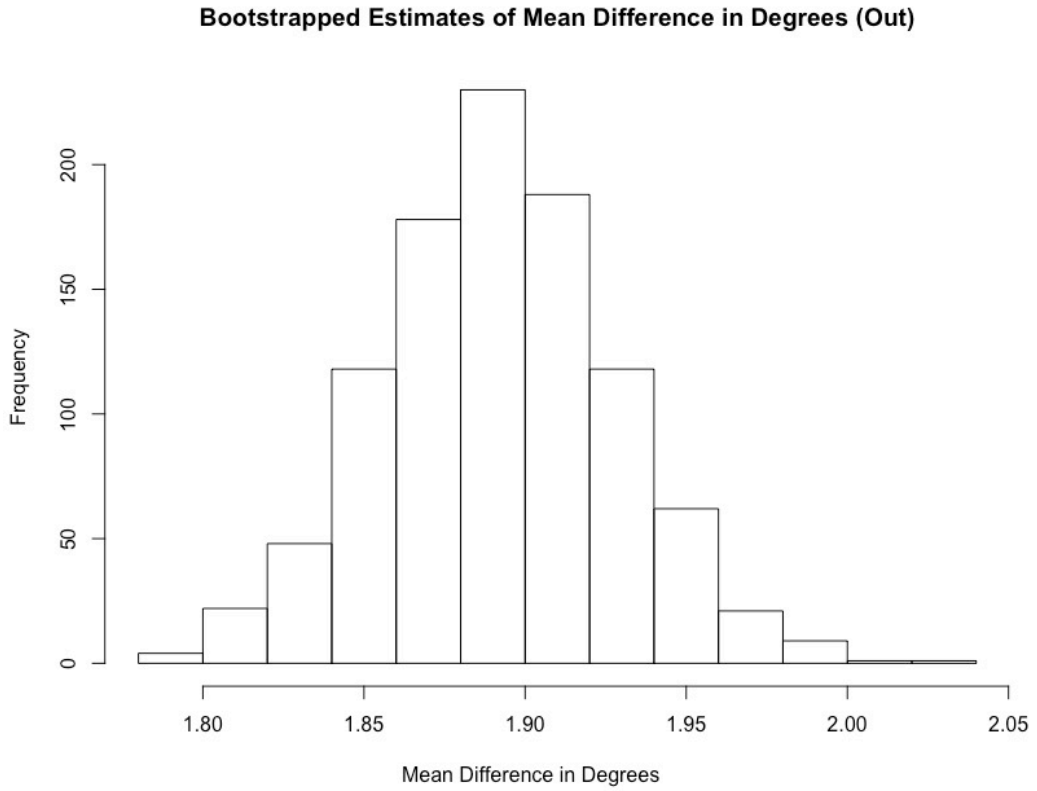



BCG

Are customers often paired with others who have **similar number** of friends, or is it typically one sided?

## **Out** Degrees

- Result: Customers typically have 1-2 more friends they interact with *than* the friends they interact with.




Insight:
*Customers shop with more friends than their friends shop with their friends. It appears "paired shopping" is not a trickle-down behavior.*

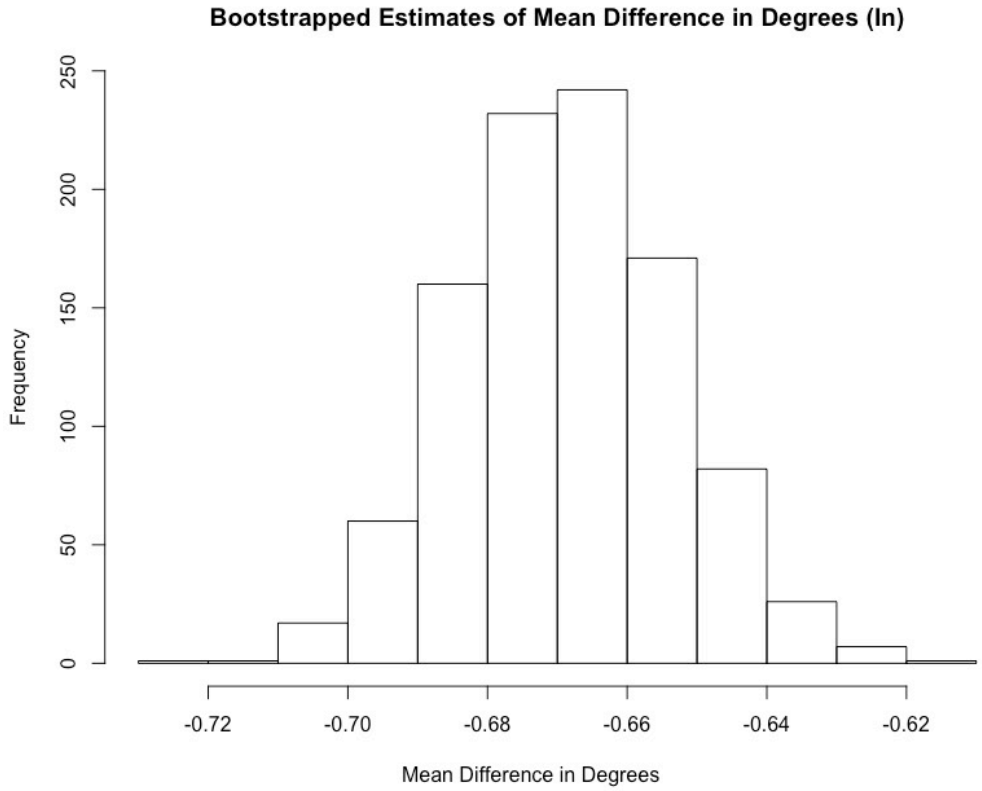
BCG

Are customers often paired with others who have **similar number** of friends, or is it typically one sided?

## **In** Degrees

- Result: Customers typically have 1 fewer friends interacting with them *than* the friends they interact with.

Insight:
*Customers are more often pulling others to shop than being pulled by others to shop.*



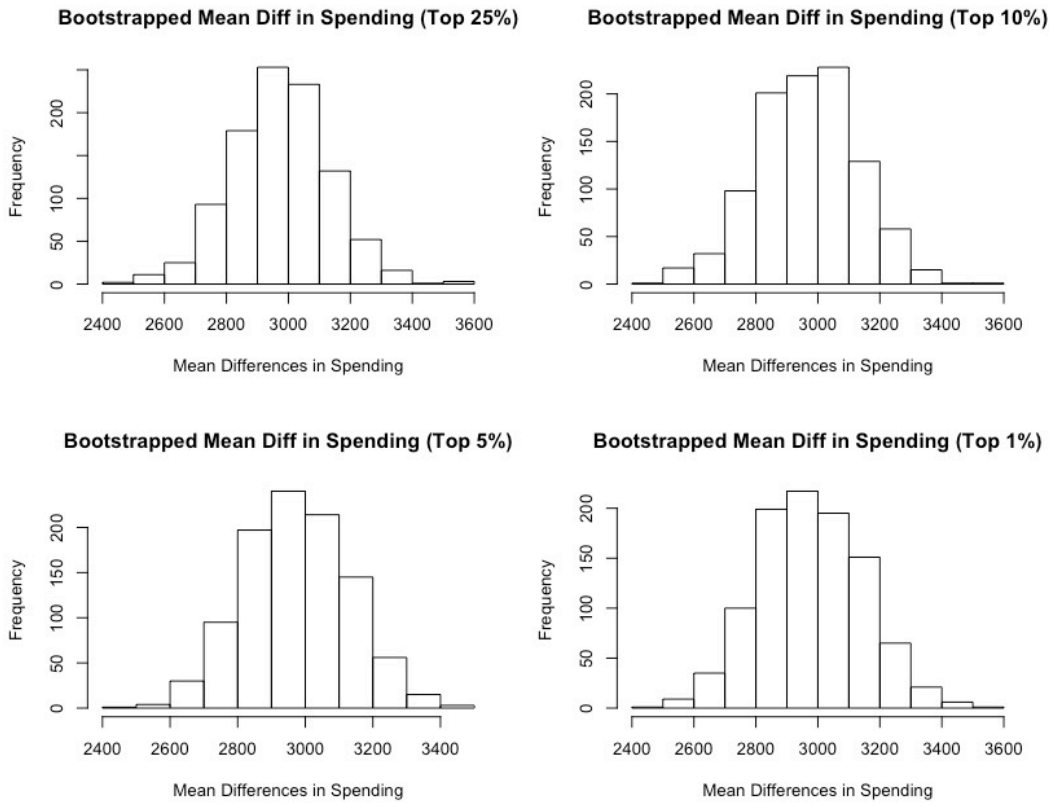

BCG

Are customers often paired with others who have **similar spending** of friends, or is it typically one sided?

## **High** Spenders

- Result: High spenders tend to spend $3000 more on average *than* the friends they shop with.

Insight:
*High spenders typically do not shop with other high spenders.*



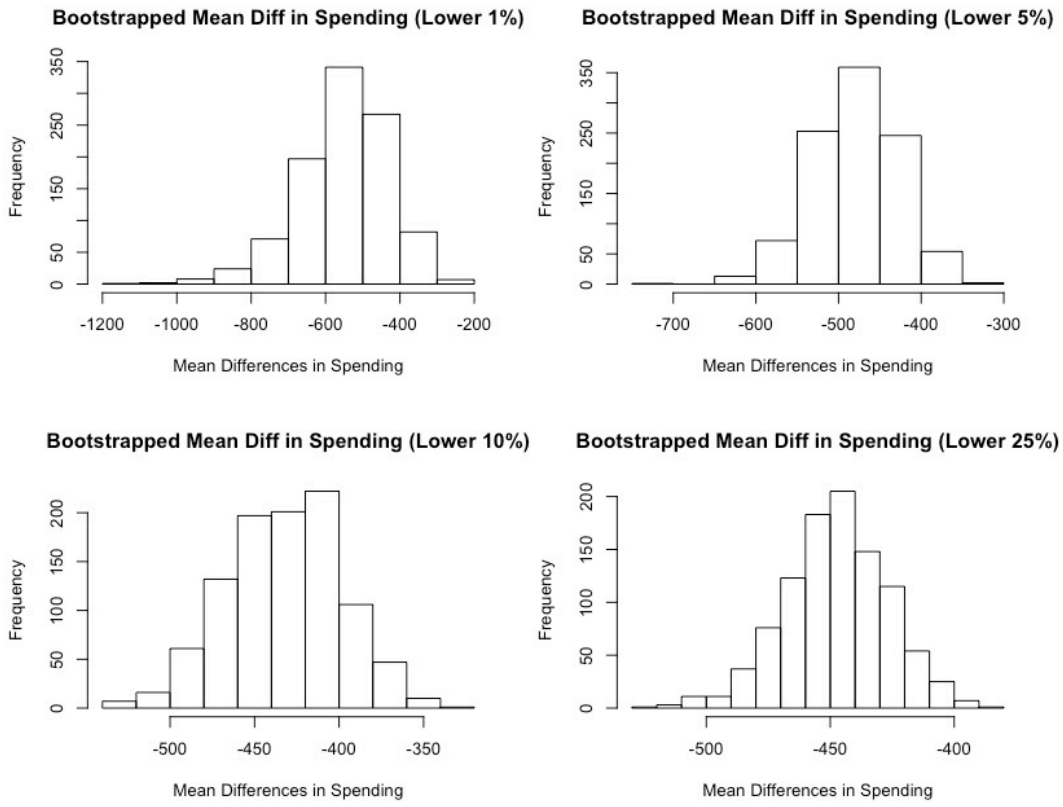

BCG

Are customers often paired with others who have **similar spending** of friends, or is it typically one sided?

## **Low** Spenders

- Result: Low spenders tend to spend $400-$600 less than the friends they shop with.

Insight:
*Low spenders typically shop with people who spend slightly more than them.*

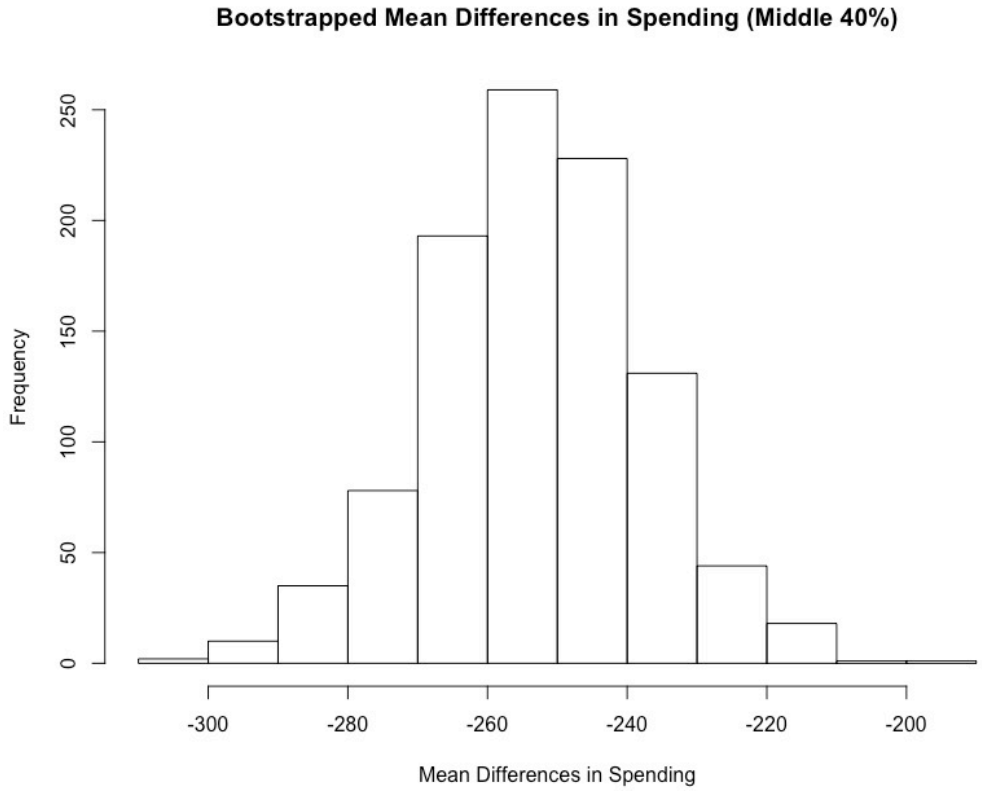


Are customers often paired with others who have **similar spending** of friends, or is it typically one sided?

## **Middle** Spenders

- Result: Middle spenders tend to spend $250 less than the friends they shop with.

Insight:
*Customers in the middle 40% of spending typically spend slightly less than the friends they shop with.*




BCG

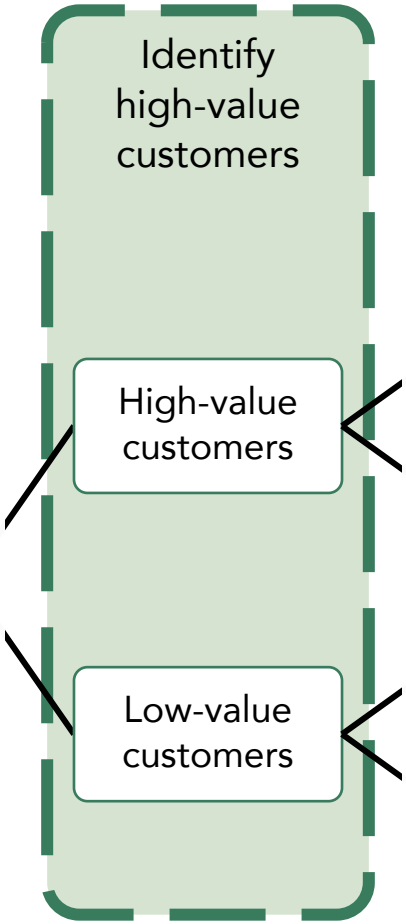
# What can we conclude on customers?

- Customers who pull their friends to shop with them rather than being pulled themselves are most likely the customers that are driving revenue.
- More often customers spend less than the friends they bring, unless these customers are very high spenders.

# Business Recommendations

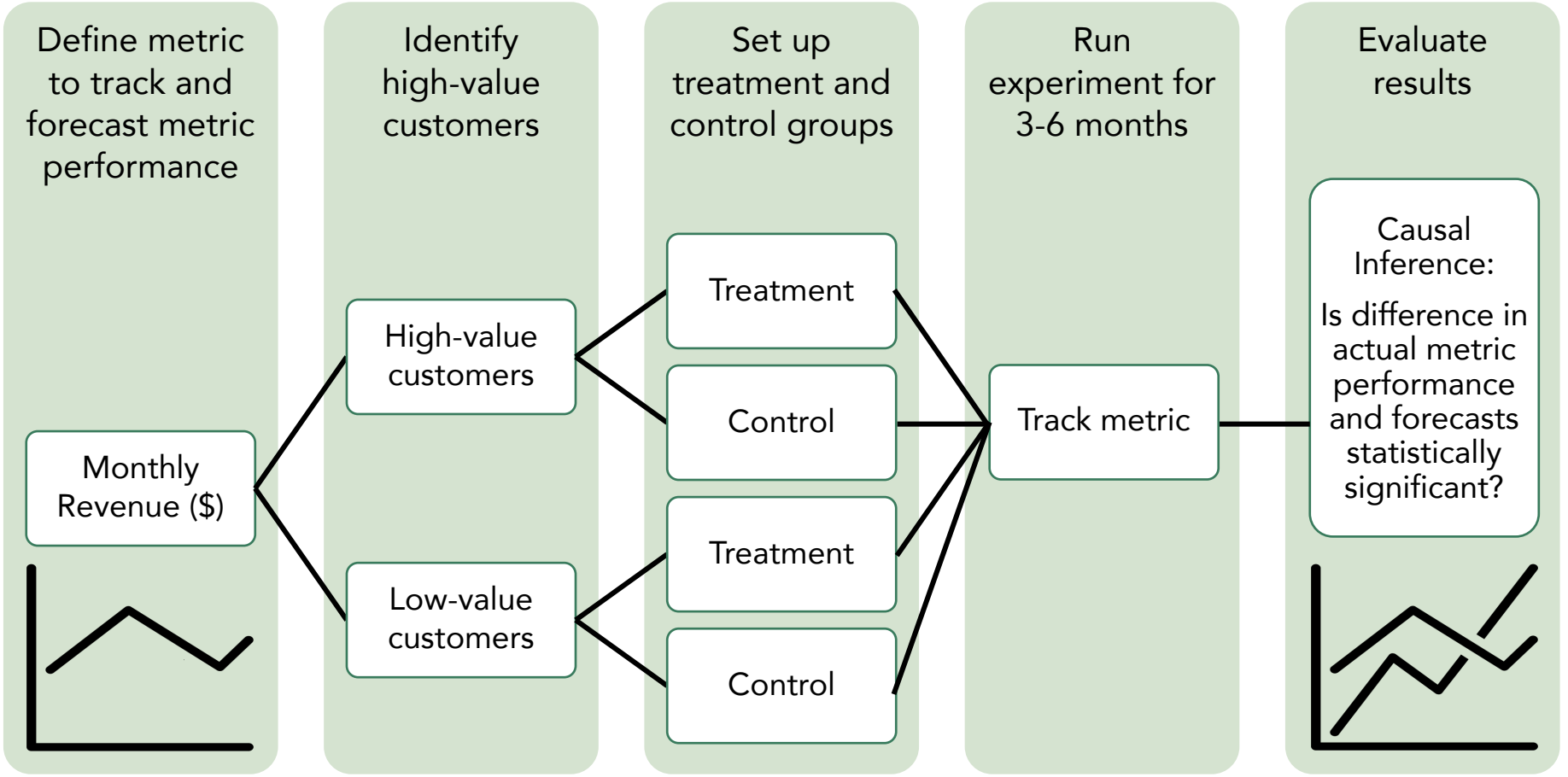
# Implement efficient spending strategy

(1) Utilize customer targeting database to identify high-value customers



# Implement efficient spending strategy

(2) Test effectiveness of new strategy with causal inference experiment



# Concluding Thoughts

# Revisiting our journey

## Primary Objective

Improve the efficiency of promotional spend via the existing loyalty program.

## Key Results Needed

- *Define* efficiency of spend
- *Quantify* efficiency of spend
- *Implement* efficient spending strategy

## Key Deliverables

- Maximize ROI by targeting high-value customers
- Utilize machine learning techniques to identify high-value customers
- Execute experiment to determine causal inference

# Revisiting our journey

## Secondary Objectives

1. *Explore* data on full-graph level and zipcode-level
2. *Deep dive* into customer relationships

## Key Results Needed

- Are there links between the size / strength of networks and certain zip codes and neighborhood-types in Seattle?
- How balanced are relationships?
  - Are customers often paired with others who have similar number of friends, or is it typically one sided?
  - Similarly for revenue – big spenders with big spenders or not?

## Key Deliverables

- There do not seem to be any links between network attributes and zip codes. However there is weak signal with certain demographic attributes.
- The subgraph size of customers is quite balanced. Spending is moderately different at lower to mid levels of spending, but extremely different at high levels of spending.

# Next steps for improved analyses

- Include data on customer response rate to loyalty program promotions
  - Why? A key factor in determining who to target with the loyalty program.
- Account for time series and seasonal effects in revenue calculations
  - Why? Eliminate confounding factors in the data.
- Collect more data on zipcode-level network behavior and demographic attributes; explore further
  - Why? Improve analysis on zipcode-level targeting

# Questions?

*Thank you for your time.*

# Appendix

Contains methodology of analysis and other bonus content!

Technical Case Interview

# Methodology: Revenue Analysis

- Full graph, Inter-zipcode (i.e. 23 subsets – one for ea. zip)
  - Pre-processing:
    - Remove observations with out-degrees < 1 or in-degrees < 1 (cannot log corresponding revenues if zeros are kept)
  - Feature engineering: 14 features from data
    - 4 features (revenue types) to be used as response variables
  - Training:
    - Train on each revenue type (response variable) one at a time
    - 5-fold cross-validation
    - Models:
      - L1-regularized OLS regression (1000 times with different seed)
        - Count most frequent features selected, average coefficients
      - Random forest (500 trees)
        - Record variable importance (% increase in MSE)
  - Most important features selected using results from both models
- Intra-zipcode
  - Pre-processing:
    - Group customers into their respective zipcodes
  - Feature engineering: 9 features from data
    - Response variable is "average spending of customers in $\text{zipcode}_i$"
  - Training:
    - Lasso regression and random forest (same as above)
  - Most important features selected using results from both models

# Methodology: Relationships Analysis

- Difference in size of friend groups of customer pairs?
  - Metrics to test:
    - Total degree (Out degree + In degree)
    - Out degree
    - In degree
  - Hypothesis testing:
    - Bootstrapped t-tests (1000x)
    - Result is distribution of mean difference in degrees
- Difference in spending of customer pairs and different segments of customer spending?
  - Setting thresholds
    - Define high-spending: Top…1%, 5%, 10%, 25%
    - Define low-spending: Bottom…1%, 5%, 10%, 25%
    - Define middle-spending: Middle…40%
  - Subset data at each threshold level
  - Bootstrapped t-tests (1000x) for differences of means in spending at each thresholds