SMM638 Mid-Term Project Description

Simone Santoni

Abstract

This document illustrates i. the business problem students are required to addres, ii. the dataset to be used to address the business problem, iii. the template that the students shall adopt to present their solution to the business problem.

1 Business problem description

Silicon is a global semiconductor company whose yearly R&D expenditures exceed USD 3B. Some Silicon's research activities concern technological opportunities that the company may exploit in the future. These opportunities depart from the core business of the company. Instead, other research activities are strongly associated with Silicon's technological and market experience. The latter concern incremental innovations in the area of microprocessors (Figure 1 shows a silicon wafer from which integrated circuit are created).

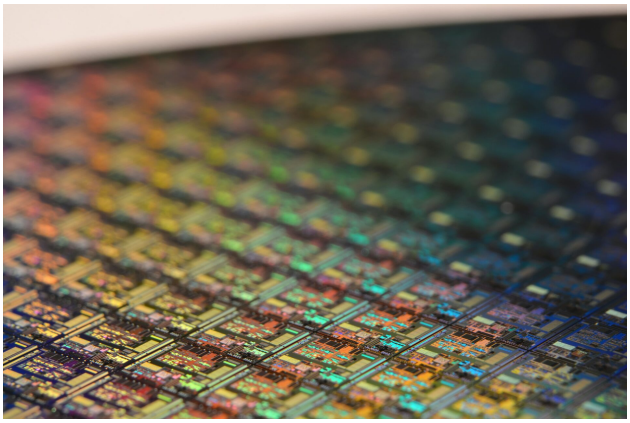

Figure 1: A silicon wafer

Silicon's management hold that there is room to improve the performance of R&D in the area of microprocessors. Particularly, the management aim to reduce the quota of abandoned/failed R&D projects and increase the economic efficiency of the R&D projects that are technically successful.

Silicon's R&D in the area of microprocessors has a project-based structure. The head of R&D allocates individual technicians and engineers to teams, which are supposed to carry out one specific project at a time. At Silicon, R&D tasks are partitioned into projects whose average duration is circa three months. Teams dissolve once they accomplish their projects or abandon a project because of technical failure.

Silicon's is your client. You are supposed to help the client improve the efficiency of R&D projects regarding microprocessors. In order to formulate your recommendations, you have been provided with data regarding a bunch of projects carried out in the first quarter of 2024 and network data concerning the knowledge sharing relationship among the technicians and engineers working in the area of microprocessors. The following sections deal with the structure and nature of these data.

2 Data

You have been provided with three datasets, which I briefly illustrate in the following sub-sections.

2.1 Team-employee affiliations

`team_employee_affiliations.csv` concerns the affiliation of the R&D personnel with the teams that have been active in the first quarter of 2024:

- The first column indicates the numeric identifier of the team
- The second column reports employees' numeric identifiers

```
import pandas as pd
teams = pd.read_csv("team_employee_affiliations.csv")
teams.head()
```

	team_id	empl_id
0	8	1
1	82	2
2	70	3
3	51	4
4	28	5

2.2 Project outcomes

`project_outcomes.csv` regards project-level outcomes, i.e., the results achieved by the individual teams:

- The first column is the team-level numeric identifier
- The second column is a binary variable discriminating between technically successful projects (`project_tech_success == 1`) and abandoned projects (`project_tech_success == 0`)
- The third column is a continuous variables expressing the number of days a team has been working on the assigned project
- The fourth column expresses the novelty (i.e., the innovativeness) of the project outcome. This variable comes from a survey circulated among the managers of the R&D department, who assessed each project against 17 likert-scale items dealing with alternative facets of the concept of novelty. Examples of items are: ''How does the project outcome depart from the R&D results Silicon achieved in the last twelve months? How does the project outcome depart from the current standard of the industry? How does the project outcome adopt atypical technical features or solutions?'' The 17 items, which all ranging from '1' to '5', were averaged to create a single indicator. Larger values of the indicator indicate more novel project outcomes.

```
outcome = pd.read_csv("project_outcomes.csv")
outcome.head()
```

	team_id	project_tech_success	project_duration	project_novelty
0	8	0	89.504485	2.946377
1	82	1	81.995252	2.863422

	team_id	project_tech_success	project_duration	project_novelty
2	70	1	90.027666	2.580469
3	51	1	90.099652	4.530003
4	28	1	87.894811	1.764565

```
outcome.describe().T
```

	count	mean	std	min	25%	50%	75%	max
team_id	101.0	51.000000	29.300171	1.000000	26.000000	51.000000	76.000000	101.000000
project_tech_success	101.0	0.772277	0.421454	0.000000	1.000000	1.000000	1.000000	1.000000
project_duration	101.0	87.298610	4.272208	76.966203	84.255601	87.822158	90.027666	100.726983
project_novelty	101.0	3.024340	0.656166	1.342613	2.555055	3.060508	3.486283	4.600945

2.3 Knowledge exchange network

The third dataset concerns the network of knowledge sharing among R&D personnel. This is a one-mode, undirected, and unweighted network. Silicon gathered these data by running a survey late December 2023. The survey asked the individual employees to name the colleagues with which they have exchanged technical and scientific knowledge in the fourth quarter of the 2023. Columns `u` and `v` contain numeric employee-level identifiers. In other words, each row of the dataframe reports a knowledge sharing tie between two employees.

```
ke = pd.read_csv(
    "knowledge_exchange_network.csv",
    sep=",",
    header=None,
    names=["u", "v"]
    )
ke.head()
```

	u	v
0	0	1
1	0	542
2	0	2
3	0	541
4	0	3

As the following cell shows, it is possible to create a NetworkX's `Graph` type object by passing a Pandas `DataFrame` to the function `nx.from_pandas_edgelist()`.

```
import networkx as nx
g = nx.from_pandas_edgelist(ke, source="u", target="v")
```

It is possible to check that `g` is undirected as follows.

```
nx.is_directed(g)
```

```
False
```

It is also possible to check that `g` is unweighted.

```
nx.is_weighted(g)
```

```
False
```

For illustrative purposes, Figure 2 pictorially depicts `g`.

```
nx.draw_kamada_kawai(g, node_size=10, node_color="lime", alpha=0.5)
```

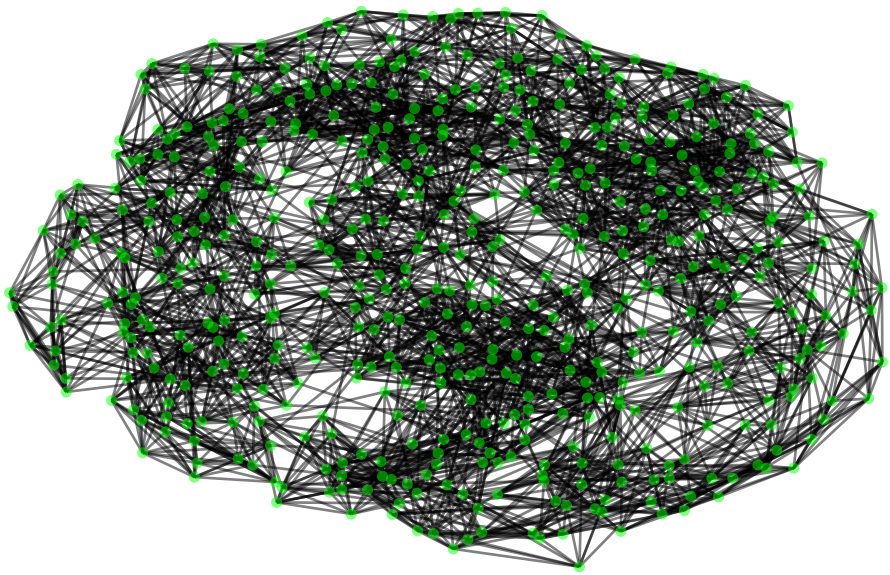

Figure 2: A visualization of Silicon's knowledge exchange network

3 Mid-term project submission template

The uploaded materials shall adhere with the following document template:

1. Data analysis procedure (max 500 words). Students may want to:
    - Give an overview of the main data analysis steps
    - Illustrate the individual steps and their rationales
2. Results (max 1,000 words). Students may want to:
    - Present a maximum of three empirical results emerging from the analysis. Incidentally, it is fine to emphasize a single result, when interesting and consequential for the formulation of your business recommendations

3. Business recommendations (1,000 word)
    - Present a maximum of three business recommendations clearly grounded in the proposed empirical results

4 Possible workflow to address the business problem

1. Consider the knowledge exchange network and compute network analysis indicators regarding density and bridging ties (which are a key features of small-world networks). You will achieve various dictionaries whose length is equal to the length of `teams`
2. Arrange these metrics in a Pandas `DataFrame`
3. Merge this Pandas `DataFrame` with the file `teams`
4. Summarise the data at the team-level so to compute, for example, the average degree across the members of the same team. To do so, you can group the data using Pandas function `groupby` along with the `aggregate` option
5. Merge the aggregate data with `outcome`
6. Run some statistical analyses highlighting the relationship between the average value of a network analysis indicator at the team level and the results achieved by the team
7. Reason on the economic and organizational significance of these results