SURVEY DESIGN

- **The Problem**
- Consider a company that sells *k* products and has a database containing the purchase histories of a large number of customers.

- The company wishes to conduct a survey, sending customized questionnaires to a particular group of *n* of its customers, to try determining which products people like overall.

Here are the guidelines for designing the survey.

- Each customer will receive questions about a certain subset of the products.
- A customer can only be asked about products that he or she has purchased.
- To make each questionnaire informative, but not too long so as to discourage participation, each customer i should be asked about a number of products between c_i and c_i'.
- Finally, to collect sufficient data about each product, there must be between p_j and p_j' distinct customers asked about each product j.

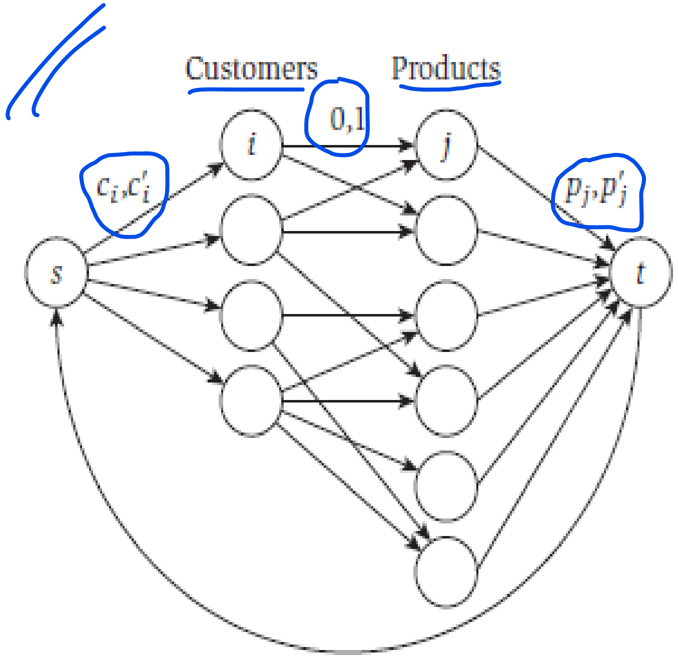
- Input to the *Survey Design Problem* consists of a bipartite graph G whose nodes are the customers and the products, and there is an edge between customer i and product j if he or she has ever purchased product j.

- for each customer $i = 1, \ldots, n$, we have limits $ci \leq ci'$ on the number of products he or she can be asked about.

- for each product $j = 1, \ldots, k$, we have limits $pj \leq pj'$ on the number of distinct customers that have to be asked about it.

- The problem is to decide if there is a way to design a questionnaire for each customer so as to satisfy all these conditions.

Designing the Algorithm



- To obtain the graph *G'* from *G*, we orient the edges of *G* from customers to products, add nodes *s* and *t* with edges *(s, i)* for each customer *i* = 1, . . . , *n*, edges *(j, t)* for each product *j* = 1, . . . , *k*, and an edge *(t, s)*.

- The flow on the edge *(s, i)* is the **number of products** included on the questionnaire for customer *i*, so this edge will have a capacity of *ci'* and a lower bound of *ci*.

- The flow on the edge *(j, t)* will correspond to the **number of customers** who were asked about product *j*, so this edge will have a capacity of *pj'* and a lower bound of *pj*.

- Each edge *(i, j)* going from a customer to a product he or she bought has capacity 1, and 0 as the lower bound. The flow carried by the edge *(t, s)* corresponds to the **overall number of questions asked**. We can give this edge a capacity of $\sum_i c_i'$ and a lower bound of $\sum_i c_i$.

Analyzing the Algorithm

- *The graph G just constructed has a feasible circulation if and only if there is a feasible way to design the survey.*

- **Proof.** The edge *(i, j)* will carry one unit of flow if customer *i* is asked about product *j* in the survey, and will carry no flow otherwise.

- The flow on the edges *(s, i)* is the number of questions asked from customer *i*.
- The flow on the edge *(j, t)* is the number of customers who were asked about product *j*, and finally.
- The flow on edge *(t, s)* is the overall number of questions asked.
- Customer *i* will be surveyed about product *j* if and only if the edge *(i, j)* carries a unit of flow.

- This flow satisfies the 0 demand, that is, there is flow conservation at every node.
- NOTE:

Circulations with Demands

- Suppose we have multiple sources and multiple sinks.
- Each sink wants to get a certain amount of flow (its **demand**).
- Each source has a certain amount of flow to give (its **supply**).
- We can represent supply as **negative demand**.

- We assume that demand and supply are perfectly matched overall. that is,

$\sum_v d_v = 0$. (d_v is the demand of the vertex v).

- Our goal is to find a flow such that everyone's demand is met (exactly), while incurring the minimum total transportation cost.