# Submodular Functions, Optimization, and Applications to Machine Learning

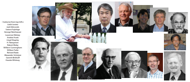
## — Spring Quarter, Lecture 1 —

http://j.ee.washington.edu/~bilmes/classes/ee596b_spring_2014/


Prof. Jeff Bilmes

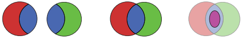
University of Washington, Seattle
Department of Electrical Engineering
http://melodi.ee.washington.edu/~bilmes

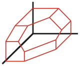

Mar 31st, 2014



$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$$

$= f(A_r) + 2f(C) + f(B_r)$ $\quad$ $= f(A_r) + f(C) + f(B_r)$ $\quad$ $= f(A \cap B)$





# Announcements

- Welcome to the class!
- Submodular Functions, Optimization, and Applications to Machine Learning, EE596B.
- Paccar 492.
- Weekly Office Hours: Wednesdays, 3:30-4:30, 10 minutes after class ends on Wednesdays.
- Class web page is at our web page (`http://j.ee.washington.edu/~bilmes/classes/ee596b_spring_2014/`)

# Directions to Paccar Hall from the EECS building

- `http://goo.gl/maps/5P3dQ`

**Suggested routes**

| | |
|---|---|
| **King Ln** | **0.4 mi, 10 mins** |
| Spokane Ln | 0.5 mi, 10 mins |
| Pierce Ln | 0.5 mi, 10 mins |

**Walking directions to Unknown road**

A **Benton Ln**

1. Head **northwest** on **Benton Ln/Grant Ln** — 443 ft
2. Turn right toward **Skagit Ln** — 236 ft
3. Turn left toward **Skagit Ln** — 0.1 mi
4. Turn right onto **Skagit Ln** — 75 ft
5. Turn left onto **King Ln**
   Take the stairs — 459 ft
6. Turn right toward **Spokane Ln** — 39 ft
7. Turn left toward **Spokane Ln** — 344 ft
8. Turn right onto **Spokane Ln** — 30 ft
9. Turn left
   Destination will be on the right — 82 ft

B **Unknown road**

# About

This course will serve as an introduction to submodular functions including methods for their optimization, and how they have been (and can be) applied in many application domains.

# Rough Outline

- Introduction to submodular functions, including definitions, real-world and contrived examples of submodular functions, properties, operations that preserve submodularity, submodular variants and special submodular functions, and computational properties.
- Background on submodular functions, including a brief overview of the theory of matroids and lattices.
- Polyhedral properties of submodular functions
- The Lovász extension of submodular functions. The Choquet integral.
- Submodular maximization algorithms under simple constraints, submodular cover problems, greedy algorithms, approximation guarantees

# Rough Outline (cont. II)

- Submodular minimization algorithms, a history of submodular minimization, including both numerical and combinatorial algorithms, computational properties of these algorithms, and descriptions of both known results and currently open problems in this area.
- Submodular flow problems, the principle partition of a submodular function and its variants.
- Constrained optimization problems with submodular functions, including maximization and minimization problems with various constraints. An overview of recent problems addressed in the community.
- Applications of submodularity in computer vision, constraint satisfaction, game theory, information theory, norms, natural language processing, graphical models, and machine learning

# Classic References

- Jack Edmonds's paper "Submodular Functions, Matroids, and Certain Polyhedra" from 1970.
- Nemhauser, Wolsey, Fisher, "A Analysis of Approximations for Maximizing Submodular Set Functions-I", 1978
- Lovász's paper, "Submodular functions and convexity", from 1983.

# Useful Books

- Fujishige, "Submodular Functions and Optimization", 2005
- Narayanan, "Submodular Functions and Electrical Networks", 1997
- Welsh, "Matroid Theory", 1975.
- Oxley, "Matroid Theory", 1992 (and 2011).
- Lawler, "Combinatorial Optimization: Networks and Matroids", 1976.
- Schrijver, "Combinatorial Optimization", 2003
- Gruenbaum, "Convex Polytopes, 2nd Ed", 2003.
- Additional readings that will be announced here.

# Recent online material (some with an ML slant)

- Previous version of this class http://j.ee.washington.edu/~bilmes/classes/ee596a_fall_2012/.
- Stefanie Jegelka & Andreas Krause's 2013 ICML tutorial http://techtalks.tv/talks/submodularity-in-machine-learning-new-directions-part-i/58125/
- NIPS, 2013 tutorial on submodularity http://melodi.ee.washington.edu/~bilmes/pgs/b2hd-bilmes2013-nips-tutorial.html and http://youtu.be/c4rBof38nKQ
- Andreas Krause's web page http://submodularity.org.
- Francis Bach's updated 2013 text. http://hal.archives-ouvertes.fr/docs/00/87/06/09/PDF/submodular_fot_revised_hal.pdf
- Tom McCormick's overview paper on submodular minimization http://people.commerce.ubc.ca/faculty/mccormick/sfmchap8a.pdf
- Georgia Tech's 2012 workshop on submodularity: http://www.arc.gatech.edu/events/arc-submodularity-workshop

# Facts about the class

- Prerequisites: ideally knowledge in probability, statistics, convex optimization, and combinatorial optimization these will be reviewed as necessary. The course is open to students in all UW departments. Any questions, please contact me.
- Text: We will be drawing from the book by Satoru Fujishige entitled "Submodular Functions and Optimization" 2nd Edition, 2005, but we will also be reading research papers that will be posted here on this web page, especially for some of the application areas.
- Grades and Assignments: Grades will be based on a combination of a final project (45%), homeworks (55%). There will be between 3-6 homeworks during the quarter.
- Final project: The final project will consist of a 4-page paper (conference style) and a final project presentation. The project must involve using/dealing mathematically with submodularity in some way or another.

# Facts about the class

- Homework/~~midterm~~ must be submitted electronically using our assignment dropbox (`https://canvas.uw.edu/courses/895956/assignments`). PDF submissions only please. Photos of neatly hand written solutions, combined into one PDF, are fine
- Lecture slides - are being prepared as we speak. I will try to have them up on the web page the night before each class. I will not only draw from the book but other sources which will be listed at the end of each set of slides.
- Slides from previous version of this class are at `http://j.ee.washington.edu/~bilmes/classes/ee596a_fall_2012/`.

# Other logistics

- Almost all equations will have numbers.

# Other logistics

- Almost all equations will have numbers.
- Equations will be numbered with lecture number, and number within lecture in the form $(\ell.j)$ where $\ell$ is the lecture number and $j$ is the $j^{\text{th}}$ equation in lecture $\ell$. For example,

$$f(A) = f(V \setminus A) \tag{1.1}$$

By the way $V \setminus A \equiv \{v \in V : v \notin A\}$ is set subtraction, sometimes written as $V - A$.

# Other logistics

- Almost all equations will have numbers.
- Equations will be numbered with lecture number, and number within lecture in the form $(\ell.j)$ where $\ell$ is the lecture number and $j$ is the $j^{\text{th}}$ equation in lecture $\ell$. For example,

$$f(A) = f(V \setminus A) \tag{1.1}$$

  By the way $V \setminus A \equiv \{v \in V : v \notin A\}$ is set subtraction, sometimes written as $V - A$.
- Theorems, Lemmas, postulates, etc. will be numbered with $(\ell.s.j)$ where $\ell$ is the lecture number, $s$ is the section number, and $j$ is the order within that section.

# Other logistics

- Almost all equations will have numbers.
- Equations will be numbered with lecture number, and number within lecture in the form $(\ell.j)$ where $\ell$ is the lecture number and $j$ is the $j^{\text{th}}$ equation in lecture $\ell$. For example,
$$f(A) = f(V \setminus A) \tag{1.1}$$
By the way $V \setminus A \equiv \{v \in V : v \notin A\}$ is set subtraction, sometimes written as $V - A$.
- Theorems, Lemmas, postulates, etc. will be numbered with $(\ell.s.j)$ where $\ell$ is the lecture number, $s$ is the section number, and $j$ is the order within that section.

**Theorem 1.1.1 (foo's theorem)**

*foo*

# Other logistics

- Almost all equations will have numbers.
- Equations will be numbered with lecture number, and number within lecture in the form $(\ell.j)$ where $\ell$ is the lecture number and $j$ is the $j^{\text{th}}$ equation in lecture $\ell$. For example,

$$f(A) = f(V \setminus A) \tag{1.1}$$

  By the way $V \setminus A \equiv \{v \in V : v \notin A\}$ is set subtraction, sometimes written as $V - A$.
- Theorems, Lemmas, postulates, etc. will be numbered with $(\ell.s.j)$ where $\ell$ is the lecture number, $s$ is the section number, and $j$ is the order within that section.

**Theorem 1.1.1 (foo's theorem)**

*foo*

- Exception to these rules is in the review sections, where theorems, equation, etc. (even if repeated) will have new reference numbers.

# Cumulative Outstanding Reading

- Read chapter 1 from Fujishige book.

# Announcements, Assignments, and Reminders

- Please do use our discussion board (https://canvas.uw.edu/courses/895956/discussion_topics) for all questions, comments, so that all will benefit from them being answered.

# Class Road Map - IT-I

- L1 (3/31): Motivation, Applications, & Basic Definitions
- L2:
- L3:
- L4:
- L5:
- L6:
- L7:
- L8:
- L9:
- L10:
- L11:
- L12:
- L13:
- L14:
- L15:
- L16:
- L17:
- L18:
- L19:
- L20:

Finals Week: June 9th-13th, 2014.

# Review

- This is where each day we will be reviewing previous lecture material.

# Submodular Definitions

**Definition 1.3.1 (submodular concave)**

A function $f : 2^V \to \mathbb{R}$ is submodular if for any $A, B \subseteq V$, we have that:

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \tag{1.2}$$

An alternate and (as we see in lecture 3) equivalent definition is:

**Definition 1.3.2 (diminishing returns)**

A function $f : 2^V \to \mathbb{R}$ is submodular if for any $A \subseteq B \subset V$, and $v \in V \setminus B$, we have that:

$$f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B) \tag{1.3}$$

This means that the incremental "value", "gain", or "cost" of $v$ decreases (diminishes) as the context in which $v$ is considered grows from $A$ to $B$.

# Sets and set functions

We are given a finite "ground" set of objects:

$V =$ 

Also given a set function $f : 2^V \to \mathbb{R}$ that valuates subsets $A \subseteq V$.

Ex: $f(V) = 6$

# Sets and set functions

Subset $A \subseteq V$ of objects:

$A = \Bigg\{$  $\Bigg\}$

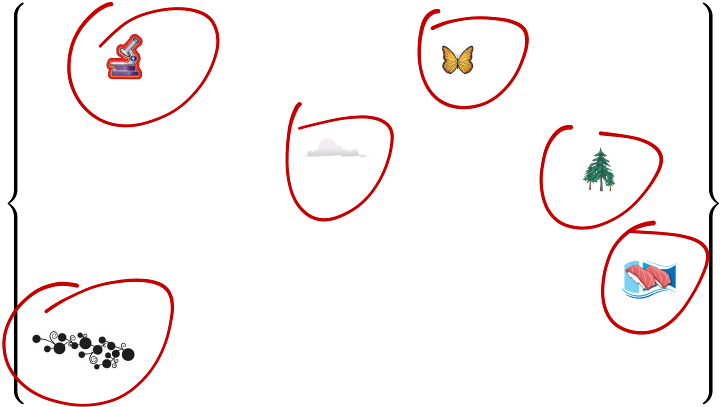
Also given a set function $f : 2^V \to \mathbb{R}$ that valuates subsets $A \subseteq V$.

Ex: $f(A) = 1$

# Sets and set functions

Subset $B \subseteq V$ of objects:

$B =$



Also given a set function $f : 2^V \to \mathbb{R}$ that valuates subsets $A \subseteq V$.
Ex: $f(B) = 6$

# Discrete Optimization Problems

- We are given a finite set of objects $V$ of size $n = |V|$.

# Discrete Optimization Problems

- We are given a finite set of objects $V$ of size $n = |V|$.
- We consider subsets of $V$. There are $2^n$ such subsets (denoted $2^V$)

# Discrete Optimization Problems

- We are given a finite set of objects $V$ of size $n = |V|$.
- We consider subsets of $V$. There are $2^n$ such subsets (denoted $2^V$)
- We have a function $f : 2^V \to \mathbb{R}$ that judges the quality (or value, or cost, or etc.) of each subset.

# Discrete Optimization Problems

- We are given a finite set of objects $V$ of size $n = |V|$.
- We consider subsets of $V$. There are $2^n$ such subsets (denoted $2^V$)
- We have a function $f : 2^V \to \mathbb{R}$ that judges the quality (or value, or cost, or etc.) of each subset.
- We may be interested only in a subset of the set of possible subsets, namely $\mathcal{S} \subseteq 2^V$. E.g., $\mathcal{S} = \{S \subseteq V : |S| \leq k\}$. The set of sets $\mathcal{S}$ might or might not itself be a function of $f$ (e.g., $\mathcal{S} = \{S \subseteq V : f(S) \leq \alpha\}$.

# Discrete Optimization Problems

- We are given a finite set of objects $V$ of size $n = |V|$.
- We consider subsets of $V$. There are $2^n$ such subsets (denoted $2^V$)
- We have a function $f : 2^V \to \mathbb{R}$ that judges the quality (or value, or cost, or etc.) of each subset.
- We may be interested only in a subset of the set of possible subsets, namely $\mathcal{S} \subseteq 2^V$. E.g., $\mathcal{S} = \{S \subseteq V : |S| \leq k\}$. The set of sets $\mathcal{S}$ might or might not itself be a function of $f$ (e.g., $\mathcal{S} = \{S \subseteq V : f(S) \leq \alpha\}$.
- A general discrete optimization problem we consider here is:

$$\begin{array}{ll} \underset{S \subseteq 2^V}{\text{maximize}} & f(S) \\ \text{subject to} & S \in \mathcal{S} \end{array} \tag{1.4}$$

# Discrete Optimization Problems

- We are given a finite set of objects $V$ of size $n = |V|$.
- We consider subsets of $V$. There are $2^n$ such subsets (denoted $2^V$)
- We have a function $f : 2^V \to \mathbb{R}$ that judges the quality (or value, or cost, or etc.) of each subset.
- We may be interested only in a subset of the set of possible subsets, namely $\mathcal{S} \subseteq 2^V$. E.g., $\mathcal{S} = \{S \subseteq V : |S| \le k\}$. The set of sets $\mathcal{S}$ might or might not itself be a function of $f$ (e.g., $\mathcal{S} = \{S \subseteq V : f(S) \le \alpha\}$.
- A general discrete optimization problem we consider here is:

$$\begin{array}{ll} \underset{S \subseteq 2^V}{\text{maximize}} & f(S) \\ \text{subject to} & S \in \mathcal{S} \end{array} \tag{1.4}$$

- Alternatively, we may minimize rather than maximize.

# Set functions are pseudo-Boolean functions

- Any set $A \subseteq V$ can be represented as a binary vector $x \in \{0,1\}^V$ (a "bit vector" representation of a set).

# Set functions are pseudo-Boolean functions

- Any set $A \subseteq V$ can be represented as a binary vector $x \in \{0,1\}^V$ (a "bit vector" representation of a set).
- The characteristic vector of a set is given by $\mathbf{1}_A \in \{0,1\}^V$ where for all $v \in V$, we have:

$$\mathbf{1}_A(v) = \begin{cases} 1 & \text{if } v \in A \\ 0 & else \end{cases} \tag{1.5}$$

# Set functions are pseudo-Boolean functions

- Any set $A \subseteq V$ can be represented as a binary vector $x \in \{0,1\}^V$ (a "bit vector" representation of a set).
- The characteristic vector of a set is given by $\mathbf{1}_A \in \{0,1\}^V$ where for all $v \in V$, we have:

$$\mathbf{1}_A(v) = \begin{cases} 1 & \text{if } v \in A \\ 0 & else \end{cases} \tag{1.5}$$

- It is sometimes useful to go back and forth between $X$ and $x(X) \triangleq \mathbf{1}_X$.

$X(x) \subseteq V$

# Set functions are pseudo-Boolean functions

- Any set $A \subseteq V$ can be represented as a binary vector $x \in \{0,1\}^V$ (a "bit vector" representation of a set).
- The characteristic vector of a set is given by $\mathbf{1}_A \in \{0,1\}^V$ where for all $v \in V$, we have:

$$\mathbf{1}_A(v) = \begin{cases} 1 & \text{if } v \in A \\ 0 & else \end{cases} \tag{1.5}$$

- It is sometimes useful to go back and forth between $X$ and $x(X) \triangleq \mathbf{1}_X$.
- $f(x) : \{0,1\}^V \to \mathbb{R}$ is a pseudo-Boolean function, and submodular functions are a special case.

# Discrete Optimization Problems

- Ignoring how complex and general this problem can be for the moment, lets consider some possible applications.
- In the rest of this section of slides, we will see many seemingly different applications that, ultimately, you will all hopefully see are strongly related to submodularity.
- We'll see, submodularity is as common and natural for discrete problems as is convexity for continuous problems.

# Example Discrete Optimization Problems

- **Combinatorial Problems**: e.g., set cover, max $k$ coverage, vertex cover, edge cover, graph cut problems.
- **Operations Research**: facility location (uncapacited)
- **Sensor placement**
- **Information:** Information gain and feature selection, information theory
- **Mathematics:** e.g., monge matrices
- **Networks**: Social networks, influence, viral marketing, information cascades, diffusion networks
- **Graphical models**: tree distributions, factors, and image segmentation
- **Diversity** and its models
- **NLP**: Natural language processing: document summarization, web search, information retrieval
- **ML**: Machine Learning: active/semi-supervised learning
- **Economics**: markets, economies of scale

# Set Cover and Maximum Coverage

- We are given a finite set $V$ of $n$ elements and a set of subsets $\mathcal{V} = \{V_1, V_2, \ldots, V_m\}$ of $m$ subsets of $V$, so that $V_i \subseteq V$ and $\bigcup_i V_i = V$.

# SET COVER and MAXIMUM COVERAGE

- We are given a finite set $V$ of $n$ elements and a set of subsets $\mathcal{V} = \{V_1, V_2, \ldots, V_m\}$ of $m$ subsets of $V$, so that $V_i \subseteq V$ and $\bigcup_i V_i = V$.
- The goal of minimum SET COVER is to choose the smallest subset $A \subseteq [m] \triangleq \{1, \ldots, m\}$ such that $\bigcup_{a \in A} V_a = V$.

# SET COVER and MAXIMUM COVERAGE

- We are given a finite set $V$ of $n$ elements and a set of subsets $\mathcal{V} = \{V_1, V_2, \dots, V_m\}$ of $m$ subsets of $V$, so that $V_i \subseteq V$ and $\bigcup_i V_i = V$.
- The goal of minimum SET COVER is to choose the smallest subset $A \subseteq [m] \triangleq \{1, \dots, m\}$ such that $\bigcup_{a \in A} V_a = V$.
- Maximum $k$ cover: The goal in MAXIMUM COVERAGE is, given an integer $k \leq m$, select $k$ subsets, say $\{a_1, a_2, \dots, a_k\}$ with $a_i \in [m]$ such that $|\bigcup_{i=1}^{k} V_{a_i}|$ is maximized.

# SET COVER and MAXIMUM COVERAGE

- We are given a finite set $V$ of $n$ elements and a set of subsets $\mathcal{V} = \{V_1, V_2, \ldots, V_m\}$ of $m$ subsets of $V$, so that $V_i \subseteq V$ and $\bigcup_i V_i = V$.
- The goal of minimum SET COVER is to choose the smallest subset $A \subseteq [m] \triangleq \{1, \ldots, m\}$ such that $\bigcup_{a \in A} V_a = V$.
- Maximum $k$ cover: The goal in MAXIMUM COVERAGE is, given an integer $k \leq m$, select $k$ subsets, say $\{a_1, a_2, \ldots, a_k\}$ with $a_i \in [m]$ such that $|\bigcup_{i=1}^k V_{a_i}|$ is maximized.
- Both SET COVER and MAXIMUM COVERAGE are well known to be NP-hard, but have a fast greedy approximation algorithm.

# Other Covers

**Definition 1.5.1 (vertex cover)**

A *vertex cover* (an "vertex-based cover of edges") in graph $G = (V, E)$ is a set $S \subseteq V(G)$ of vertices such that every edge in $G$ is incident to at least one vertex in $S$.

- Let $I(S)$ be the number of edges incident to vertex set $S$. Then we wish to find the smallest set $S \subseteq V$ subject to $I(S) = |E|$.

**Definition 1.5.2 (edge cover)**

A *edge cover* (an "edge-based cover of vertices") in graph $G = (V, E)$ is a set $F \subseteq E(G)$ of edges such that every vertex in $G$ is incident to at least one edge in $F$.

- Let $|V|(F)$ be the number of vertices incident to edge set $F$. Then we wish to find the smallest set $F \subseteq E$ subject to $|V|(F) = |V|$.

# Graph Cut Problems

- MINIMUM CUT: Given a graph $G = (V, E)$, find a set of vertices $S \subseteq V$ that minimize the cut (set of edges) between $S$ and $V \setminus S$.

# Graph Cut Problems

- MINIMUM CUT: Given a graph $G = (V, E)$, find a set of vertices $S \subseteq V$ that minimize the cut (set of edges) between $S$ and $V \setminus S$.
- MAXIMUM CUT: Given a graph $G = (V, E)$, find a set of vertices $S \subseteq V$ that minimize the cut (set of edges) between $S$ and $V \setminus S$.

# Graph Cut Problems

- MINIMUM CUT: Given a graph $G = (V, E)$, find a set of vertices $S \subseteq V$ that minimize the cut (set of edges) between $S$ and $V \setminus S$.
- MAXIMUM CUT: Given a graph $G = (V, E)$, find a set of vertices $S \subseteq V$ that minimize the cut (set of edges) between $S$ and $V \setminus S$.
- Let $f : 2^V \to \mathbb{R}_+$ be the cut function, namely for any given set of nodes $X \subseteq V$, $f(X)$ measures the number of edges between nodes $X$ and $V \setminus X$.

# Graph Cut Problems

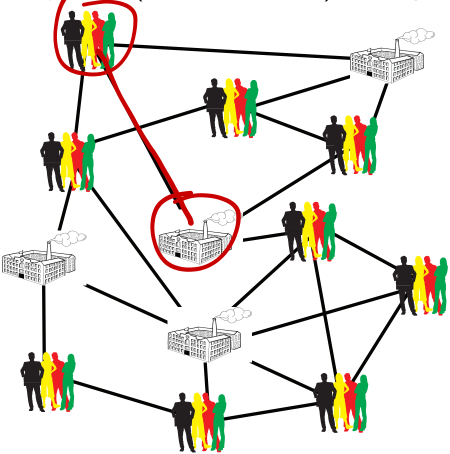
- MINIMUM CUT: Given a graph $G = (V, E)$, find a set of vertices $S \subseteq V$ that minimize the cut (set of edges) between $S$ and $V \setminus S$.
- MAXIMUM CUT: Given a graph $G = (V, E)$, find a set of vertices $S \subseteq V$ that minimize the cut (set of edges) between $S$ and $V \setminus S$.
- Let $f : 2^V \to \mathbb{R}_+$ be the cut function, namely for any given set of nodes $X \subseteq V$, $f(X)$ measures the number of edges between nodes $X$ and $V \setminus X$.
- Weighted versions, where rather than count, we sum the (non-negative) weights of the edges of a cut.

# Graph Cut Problems

- MINIMUM CUT: Given a graph $G = (V, E)$, find a set of vertices $S \subseteq V$ that minimize the cut (set of edges) between $S$ and $V \setminus S$.
- MAXIMUM CUT: Given a graph $G = (V, E)$, find a set of vertices $S \subseteq V$ that minimize the cut (set of edges) between $S$ and $V \setminus S$.
- Let $f : 2^V \to \mathbb{R}_+$ be the cut function, namely for any given set of nodes $X \subseteq V$, $f(X)$ measures the number of edges between nodes $X$ and $V \setminus X$.
- Weighted versions, where rather than count, we sum the (non-negative) weights of the edges of a cut.
- Many examples of this, we will see more later.

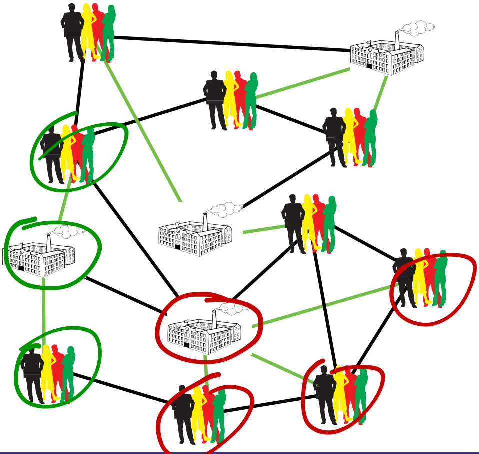
# Facility/Plant Location (uncapacitated)

- Core problem in operations research and strong early motivation for submodular functions.
- Goal: as efficiently as possible, place "facilities" (factories) at certain locations to satisfy sites (at all locations) having various demands.



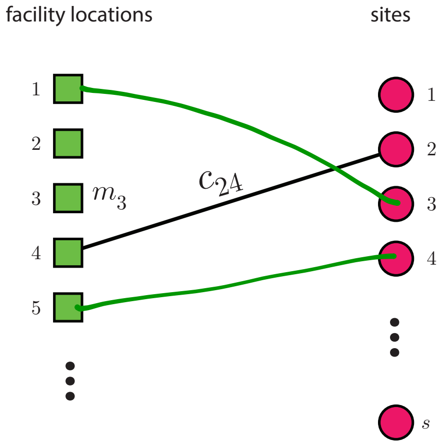
# Facility/Plant Location (uncapacitated)

- Core problem in operations research and strong early motivation for submodular functions.
- Goal: as efficiently as possible, place "facilities" (factories) at certain locations to satisfy sites (at all locations) having various demands.



# Facility/Plant Location (uncapacitated)

- Core problem in operations research and strong early motivation for submodular functions.
- Goal: as efficiently as possible, place "facilities" (factories) at certain locations to satisfy sites (at all locations) having various demands.



# Facility/Plant Location (uncapacitated)

- Let $F = \{1, \ldots, f\}$ be a set of possible factory/plant locations for facilities to be built.
- $S = \{1, \ldots, s\}$ is a set of sites needing to be serviced (e.g., cities, clients).
- Let $c_{ij}$ be the "benefit" (e.g., $1/c_{ij}$ is the cost) of servicing site $i$ with facility location $j$.
- Let $m_j$ be the benefit (e.g., either $1/m_j$ is the cost or $-m_j$ is the cost) to build a plant at location $j$.
- Each site needs to be serviced by only one plant but no less than one.
- Define $f(A)$ as the "delivery benefit" plus "construction benefit" when the locations $A \subseteq F$ are to be constructed.
- We can define $f(A) = \sum_{j \in A} m_j + \sum_{i \in F} \max_{j \in A} c_{ij}$.
- Goal is to find a set $A$ that maximizes $f(A)$ (the benefit) placing a bound on the number of plants $A$ (e.g., $|A| \leq k$).

# Sensor Placement

- Given an environment, there is a set $V$ of candidate locations for placement of a sensor (e.g., temperature, gas, audio, video, bacteria or other environmental contaminant, etc.).

# Sensor Placement

- Given an environment, there is a set $V$ of candidate locations for placement of a sensor (e.g., temperature, gas, audio, video, bacteria or other environmental contaminant, etc.).
- We have a function $f(S)$ that measures the "coverage" of any given set $S$ of sensor placement decisions. Then $f(V)$ is maximum possible coverage.

# Sensor Placement

- Given an environment, there is a set $V$ of candidate locations for placement of a sensor (e.g., temperature, gas, audio, video, bacteria or other environmental contaminant, etc.).
- We have a function $f(S)$ that measures the "coverage" of any given set $S$ of sensor placement decisions. Then $f(V)$ is maximum possible coverage.
- One possible goal: choose smallest set $S$ such that $f(S) = \alpha f(V)$ with $0 < \alpha \leq 1$.

# Sensor Placement

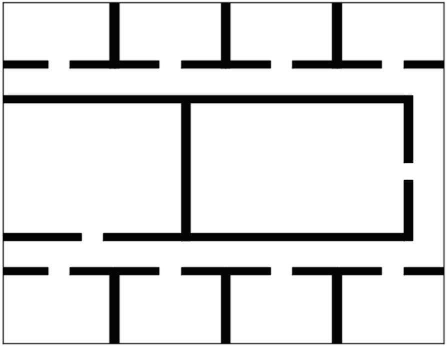
- Given an environment, there is a set $V$ of candidate locations for placement of a sensor (e.g., temperature, gas, audio, video, bacteria or other environmental contaminant, etc.).
- We have a function $f(S)$ that measures the "coverage" of any given set $S$ of sensor placement decisions. Then $f(V)$ is maximum possible coverage.
- One possible goal: choose smallest set $S$ such that $f(S) = \alpha f(V)$ with $0 < \alpha \leq 1$.
- Another possible goal: choose size at most $k$ set $S$ such that $f(S)$ is maximized.

# Sensor Placement

- Given an environment, there is a set $V$ of candidate locations for placement of a sensor (e.g., temperature, gas, audio, video, bacteria or other environmental contaminant, etc.).
- We have a function $f(S)$ that measures the "coverage" of any given set $S$ of sensor placement decisions. Then $f(V)$ is maximum possible coverage.
- One possible goal: choose smallest set $S$ such that $f(S) = \alpha f(V)$ with $0 < \alpha \leq 1$.
- Another possible goal: choose size at most $k$ set $S$ such that $f(S)$ is maximized.
- Environment could be a floor of a building, water network, monitored ecological preservation.

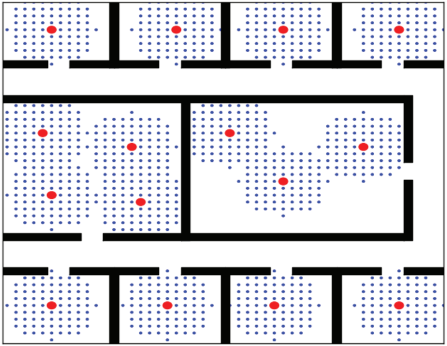
# Sensor Placement within Buildings

- An example of a room layout. Should be possible to determine temperature at all points in the room. Sensors cannot sense beyond wall (thick black line) boundaries.



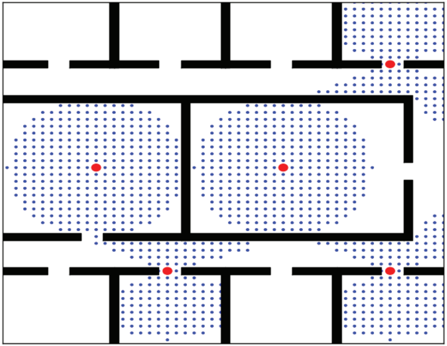
# Sensor Placement within Buildings

- Example sensor placement using small range cheap sensors (located at red dots).



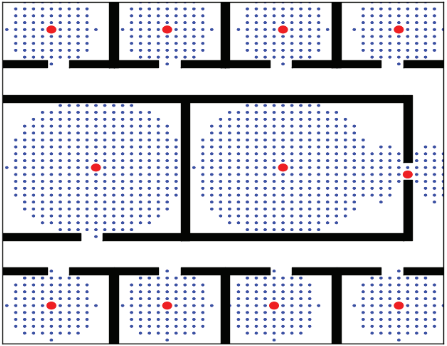
# Sensor Placement within Buildings

- Example sensor placement using longer range expensive sensors (located at red dots).



# Sensor Placement within Buildings

- Example sensor placement using mixed range sensors (located at red dots).



# Information Gain and Feature Selection

- Task: pattern recognition based on (at most) features $X_V$ to predict random variable $Y$. True model is $p(Y, X_V)$, where $V$ is a finite set of feature indices.

# Information Gain and Feature Selection

- Task: pattern recognition based on (at most) features $X_V$ to predict random variable $Y$. True model is $p(Y, X_V)$, where $V$ is a finite set of feature indices.
- Given subset of features $A \subseteq V$, prediction based on $p(y|x_A)$.

# Information Gain and Feature Selection

- Task: pattern recognition based on (at most) features $X_V$ to predict random variable $Y$. True model is $p(Y, X_V)$, where $V$ is a finite set of feature indices.
- Given subset of features $A \subseteq V$, prediction based on $p(y|x_A)$.
- Goal: choose the smallest set of features that retains accuracy.

# Information Gain and Feature Selection

- Task: pattern recognition based on (at most) features $X_V$ to predict random variable $Y$. True model is $p(Y, X_V)$, where $V$ is a finite set of feature indices.
- Given subset of features $A \subseteq V$, prediction based on $p(y|x_A)$.
- Goal: choose the smallest set of features that retains accuracy.
- Information gain is defined as:

$$f(A) = I(Y; X_A) = H(Y) - H(Y|X_A) = H(X_A) - H(X_A|Y) \tag{1.6}$$

# Information Gain and Feature Selection

- Task: pattern recognition based on (at most) features $X_V$ to predict random variable $Y$. True model is $p(Y, X_V)$, where $V$ is a finite set of feature indices.
- Given subset of features $A \subseteq V$, prediction based on $p(y|x_A)$.
- Goal: choose the smallest set of features that retains accuracy.
- Information gain is defined as:

$$f(A) = I(Y; X_A) = H(Y) - H(Y|X_A) = H(X_A) - H(X_A|Y) \tag{1.6}$$

- Goal is to find a subset $A$ of size $k$ that has high information gain.

# Information Gain and Feature Selection

- Task: pattern recognition based on (at most) features $X_V$ to predict random variable $Y$. True model is $p(Y, X_V)$, where $V$ is a finite set of feature indices.
- Given subset of features $A \subseteq V$, prediction based on $p(y|x_A)$.
- Goal: choose the smallest set of features that retains accuracy.
- Information gain is defined as:

$$f(A) = I(Y; X_A) = H(Y) - H(Y|X_A) = H(X_A) - H(X_A|Y) \tag{1.6}$$

- Goal is to find a subset $A$ of size $k$ that has high information gain.
- Applicable not only in pattern recognition, but in the sensor coverage problem as well, where $Y$ is whatever question we wish to ask about the room.

# Information Theory: Block Coding

- Given a set of random variables $\{X_i\}_{i \in V}$ indexed by set $V$, how do we partition them so that we can best block-code them within each block.

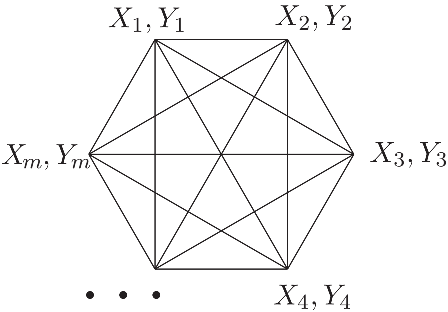
# Information Theory: Block Coding

- Given a set of random variables $\{X_i\}_{i \in V}$ indexed by set $V$, how do we partition them so that we can best block-code them within each block.
- I.e., how do we form $S \subseteq V$ such that $I(X_S; X_{V \setminus S})$ is as small as possible, where $I(X_A; X_B)$ is the mutual information between random variables $X_A$ and $X_B$, i.e.,

$$I(X_A; X_B) = H(X_A) + H(X_B) - H(X_A, X_B) \tag{1.7}$$

and $H(X_A) = -\sum_{x_A} p(x_A) \log p(x_A)$ is the joint entropy of the set $X_A$ of random variables.

# Information Theory: Networks Communication




- A network of senders/receivers
- Each sender $X_i$ is trying to communicate simultaneously with each receiver $Y_i$ (i.e., for all $i$, $X_i$ is sending to $\{Y_i\}_i$
- The $X_i$ are not necessarily independent.

- Communication rates from $i$ to $j$ are $R^{(i\to j)}$ to send message $W^{(i\to j)} \in \left\{1, 2, \ldots, 2^{nR^{(i\to j)}}\right\}$.
- Goal: necessary and sufficient conditions for achievability as we've done for other channels.
- I.e., can we find functions $f$ such that any rates must satisfy

$$\forall S \subseteq V, \sum_{i\in S, j\in V\setminus S} R^{(i\to j)} \leq f(S) \tag{1.8}$$

# Monge Matrices

- $m \times n$ matrices $C = [c_{ij}]_{ij}$ are called Monge matrices if they satisfy the Monge property, namely:

$$c_{ij} + c_{rs} \leq c_{is} + c_{rj} \tag{1.9}$$

for all $1 \leq i < r \leq m$ and $1 \leq j < s \leq n$.

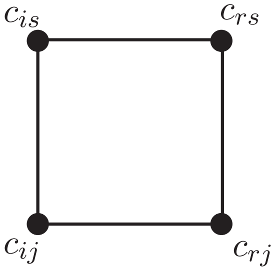
# Monge Matrices

- $m \times n$ matrices $C = [c_{ij}]_{ij}$ are called Monge matrices if they satisfy the Monge property, namely:

$$c_{ij} + c_{rs} \leq c_{is} + c_{rj} \tag{1.9}$$

for all $1 \leq i < r \leq m$ and $1 \leq j < s \leq n$.

- Consider four elements of the matrix:



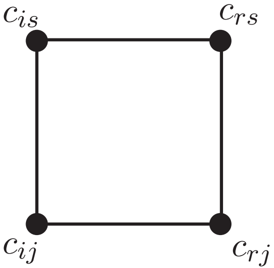
# Monge Matrices

- $m \times n$ matrices $C = [c_{ij}]_{ij}$ are called Monge matrices if they satisfy the Monge property, namely:

$$c_{ij} + c_{rs} \leq c_{is} + c_{rj} \tag{1.9}$$

for all $1 \leq i < r \leq m$ and $1 \leq j < s \leq n$.

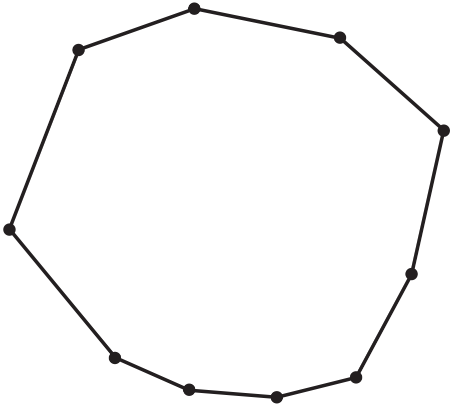
- Consider four elements of the matrix:




- Useful for speeding up certain dynamic programming problems.

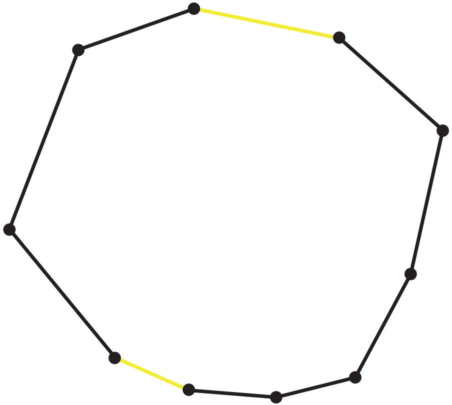
# Monge Matrices

- Can generate a Monge matrix from a convex polygon - delete two segments, then separately number vertices on each chain. Distances $c_{ij}$ satisfy Monge property (or quadrangle inequality).



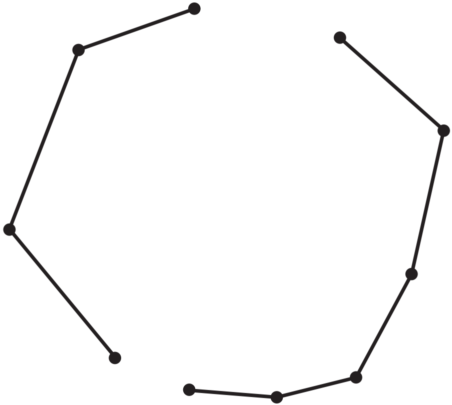
# Monge Matrices

- Can generate a Monge matrix from a convex polygon - delete two segments, then separately number vertices on each chain. Distances $c_{ij}$ satisfy Monge property (or quadrangle inequality).



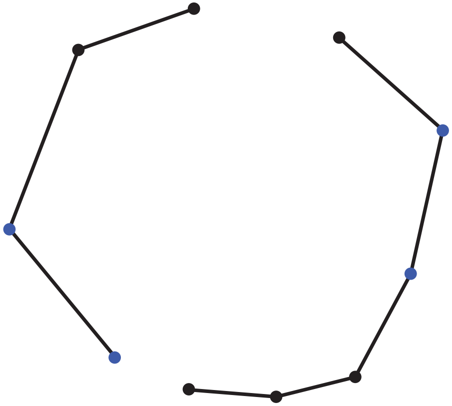
# Monge Matrices

- Can generate a Monge matrix from a convex polygon - delete two segments, then separately number vertices on each chain. Distances $c_{ij}$ satisfy Monge property (or quadrangle inequality).



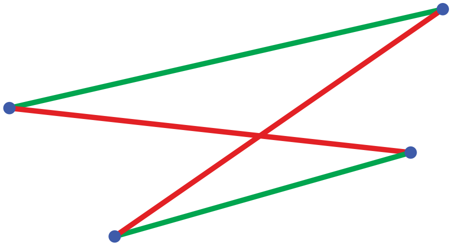
# Monge Matrices

- Can generate a Monge matrix from a convex polygon - delete two segments, then separately number vertices on each chain. Distances $c_{ij}$ satisfy Monge property (or quadrangle inequality).



# Monge Matrices

- Can generate a Monge matrix from a convex polygon - delete two segments, then separately number vertices on each chain. Distances $c_{ij}$ satisfy Monge property (or quadrangle inequality).



# A model of Influence in Social Networks

- Given a graph $G = (V, E)$, each $v \in V$ corresponds to a person, to each $v$ we have an activation function $f_v : 2^V \to [0, 1]$ dependent only on its neighbors. I.e., $f_v(A) = f_v(A \cap \Gamma(v))$.

# A model of Influence in Social Networks

- Given a graph $G = (V, E)$, each $v \in V$ corresponds to a person, to each $v$ we have an activation function $f_v : 2^V \to [0, 1]$ dependent only on its neighbors. I.e., $f_v(A) = f_v(A \cap \Gamma(v))$.
- Goal - Viral Marketing: find a small subset $S \subseteq V$ of individuals to directly influence, and thus indirectly influence the greatest number of possible other individuals (via the social network $G$).

# A model of Influence in Social Networks

- Given a graph $G = (V, E)$, each $v \in V$ corresponds to a person, to each $v$ we have an activation function $f_v : 2^V \to [0, 1]$ dependent only on its neighbors. I.e., $f_v(A) = f_v(A \cap \Gamma(v))$.
- Goal - Viral Marketing: find a small subset $S \subseteq V$ of individuals to directly influence, and thus indirectly influence the greatest number of possible other individuals (via the social network $G$).
- We define a function $f : 2^V \to \mathbb{Z}^+$ that models the ultimate influence of an initial set $S$ of nodes based on the following iterative process: At each step, a given set of nodes $S$ are activated, and we activate new nodes $v \in V \setminus S$ if $f_v(S) \geq U[0, 1]$ (where $U[0, 1]$ is a uniform random number between 0 and 1).

# The value of a friend

- Let $V$ be a group of individuals. How valuable to you is a given friend $v \in V$?

# The value of a friend

- Let $V$ be a group of individuals. How valuable to you is a given friend $v \in V$?
- It depends on how many friends you have.

# The value of a friend

- Let $V$ be a group of individuals. How valuable to you is a given friend $v \in V$?
- It depends on how many friends you have.
- Given a group of friends $S \subseteq V$, can you valuate them with a function $f(S)$ an how?

# The value of a friend

- Let $V$ be a group of individuals. How valuable to you is a given friend $v \in V$?
- It depends on how many friends you have.
- Given a group of friends $S \subseteq V$, can you valuate them with a function $f(S)$ an how?
- Let $f(S)$ be the value of the set of friends $S$. Is submodular or supermodular a good model?

# Information Cascades, Diffusion Networks

- How to model flow of information from source to the point it reaches users — information used in its common sense (like news events).

# Information Cascades, Diffusion Networks

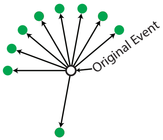
- How to model flow of information from source to the point it reaches users — information used in its common sense (like news events).
- How to find the most influential sources, the ones that often set off cascades, which are like large "waves" of information flow?

# Information Cascades, Diffusion Networks

- How to model flow of information from source to the point it reaches users — information used in its common sense (like news events).
- How to find the most influential sources, the ones that often set off cascades, which are like large “waves” of information flow?
- There might be only one seed source (shown below) or many.

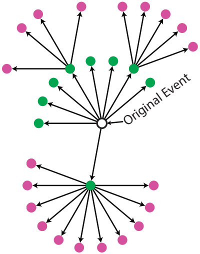
# Information Cascades, Diffusion Networks

- How to model flow of information from source to the point it reaches users — information used in its common sense (like news events).
- How to find the most influential sources, the ones that often set off cascades, which are like large "waves" of information flow?
- There might be only one seed source (shown below) or many.



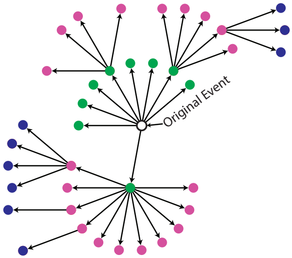
# Information Cascades, Diffusion Networks

- How to model flow of information from source to the point it reaches users — information used in its common sense (like news events).
- How to find the most influential sources, the ones that often set off cascades, which are like large "waves" of information flow?
- There might be only one seed source (shown below) or many.



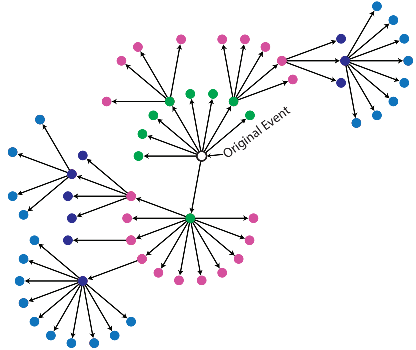
# Information Cascades, Diffusion Networks

- How to model flow of information from source to the point it reaches users — information used in its common sense (like news events).
- How to find the most influential sources, the ones that often set off cascades, which are like large "waves" of information flow?
- There might be only one seed source (shown below) or many.



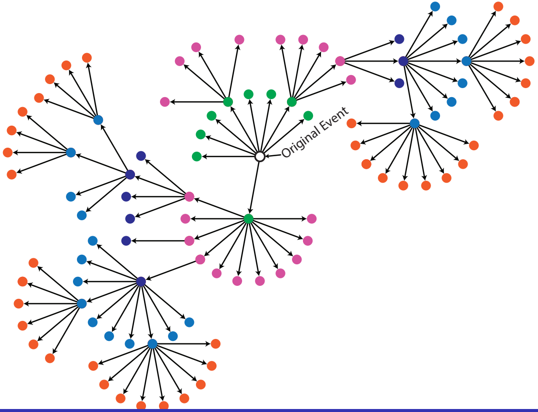
# Information Cascades, Diffusion Networks

- How to model flow of information from source to the point it reaches users — information used in its common sense (like news events).
- How to find the most influential sources, the ones that often set off cascades, which are like large "waves" of information flow?
- There might be only one seed source (shown below) or many.



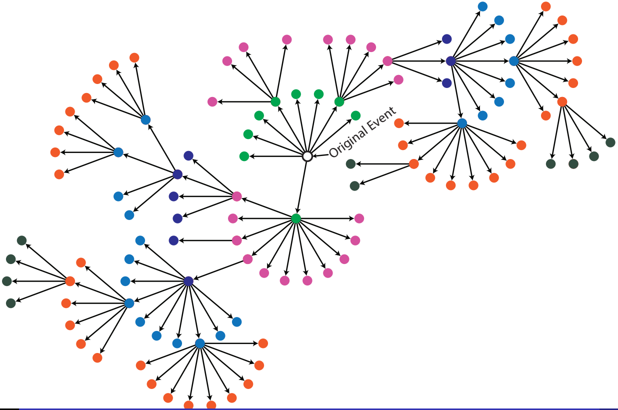
# Information Cascades, Diffusion Networks

- How to model flow of information from source to the point it reaches users — information used in its common sense (like news events).
- How to find the most influential sources, the ones that often set off cascades, which are like large "waves" of information flow?
- There might be only one seed source (shown below) or many.



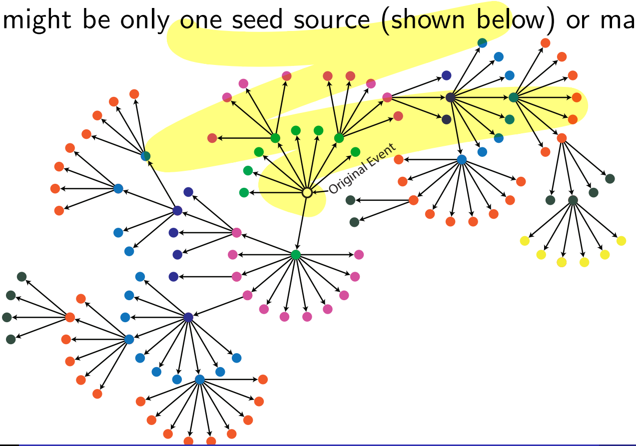
# Information Cascades, Diffusion Networks

- How to model flow of information from source to the point it reaches users — information used in its common sense (like news events).
- How to find the most influential sources, the ones that often set off cascades, which are like large "waves" of information flow?
- There might be only one seed source (shown below) or many.



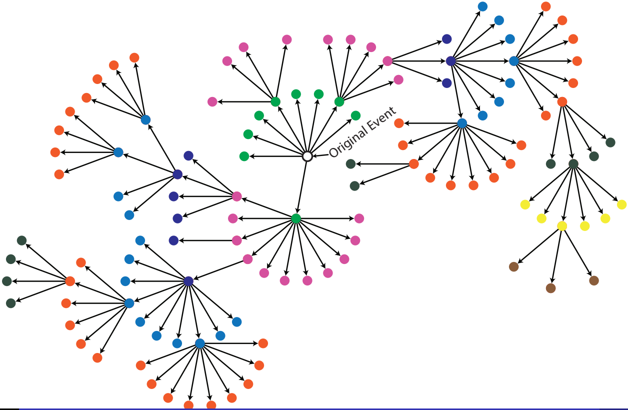
# Information Cascades, Diffusion Networks

- How to model flow of information from source to the point it reaches users — information used in its common sense (like news events).
- How to find the most influential sources, the ones that often set off cascades, which are like large "waves" of information flow?
- There might be only one seed source (shown below) or many.



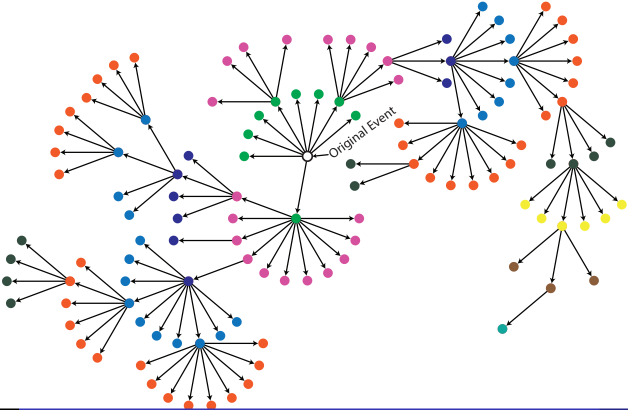
# Information Cascades, Diffusion Networks

- How to model flow of information from source to the point it reaches users — information used in its common sense (like news events).
- How to find the most influential sources, the ones that often set off cascades, which are like large "waves" of information flow?
- There might be only one seed source (shown below) or many.



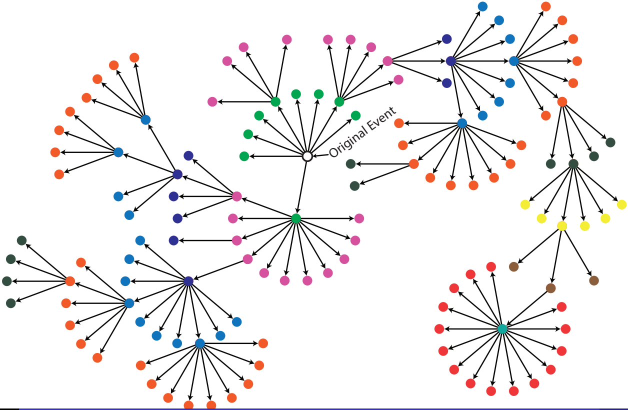
# Information Cascades, Diffusion Networks

- How to model flow of information from source to the point it reaches users — information used in its common sense (like news events).
- How to find the most influential sources, the ones that often set off cascades, which are like large "waves" of information flow?
- There might be only one seed source (shown below) or many.



# Information Cascades, Diffusion Networks

- How to model flow of information from source to the point it reaches users — information used in its common sense (like news events).
- How to find the most influential sources, the ones that often set off cascades, which are like large “waves” of information flow?
- There might be only one seed source (shown below) or many.



# Diffusion Networks

- Information propagation: when blogs or news stories break, and creates an information cascade over multiple other blogs/newspapers/magazines.
- Viral marketing: What is the pattern of trendsetters that cause an individual to purchase a product?
- Epidemiology: who got sick from whom, and what is the network of such links?
- How can we infer the connectivity of a network (of memes, purchase decisions, virusus, etc.) based only on diffusion traces (the time that each node is "infected")? How to find the most likely tree?

# Graphical Models: Tree Distributions

- Family of probability distributions $p : \{0,1\}^V \to [0,1]$:

$$p(x) = \frac{1}{Z}\exp(f(x)) \tag{1.10}$$

# Graphical Models: Tree Distributions

- Family of probability distributions $p : \{0,1\}^V \to [0,1]$:

$$p(x) = \frac{1}{Z}\exp(f(x)) \tag{1.10}$$

- Given a graphical model $G = (V, E)$ and a family of probability distributions $p \in \mathcal{F}(G, \mathcal{M})$ that factor w.r.t. that distribution. I.e., $f(x) = \sum_{c \in \mathcal{C}} f_c(x_c)$ where $\mathcal{C}$ are a set of cliques.

# Graphical Models: Tree Distributions

- Family of probability distributions $p : \{0,1\}^V \to [0,1]$:

$$p(x) = \frac{1}{Z} \exp(f(x)) \tag{1.10}$$

- Given a graphical model $G = (V, E)$ and a family of probability distributions $p \in \mathcal{F}(G, \mathcal{M})$ that factor w.r.t. that distribution. I.e., $f(x) = \sum_{c \in \mathcal{C}} f_c(x_c)$ where $\mathcal{C}$ are a set of cliques.
- Find the closest distribution $p_t$ to $p$ subject to $p_t$ factoring w.r.t. some tree $T = (V, F)$, i.e., $p_t \in \mathcal{F}(T, \mathcal{M})$.

# Graphical Models: Tree Distributions

- Family of probability distributions $p : \{0,1\}^V \to [0,1]$:

$$p(x) = \frac{1}{Z} \exp(f(x)) \tag{1.10}$$

- Given a graphical model $G = (V, E)$ and a family of probability distributions $p \in \mathcal{F}(G, \mathcal{M})$ that factor w.r.t. that distribution. I.e., $f(x) = \sum_{c \in \mathcal{C}} f_c(x_c)$ where $\mathcal{C}$ are a set of cliques.
- Find the closest distribution $p_t$ to $p$ subject to $p_t$ factoring w.r.t. some tree $T = (V, F)$, i.e., $p_t \in \mathcal{F}(T, \mathcal{M})$.
- I.e., optimization problem

$$\begin{aligned} \underset{p_t \in \mathcal{F}(G,\mathcal{M})}{\text{minimize}} \quad & D(p||p_t) \\ \text{subject to} \quad & p_t \in \mathcal{F}(T, \mathcal{M}). \\ & T = (V, F) \text{ is a tree} \end{aligned} \tag{1.11}$$

# Graphical Models: Tree Distributions

- Family of probability distributions $p : \{0,1\}^V \to [0,1]$:

$$p(x) = \frac{1}{Z}\exp(f(x)) \tag{1.10}$$

- Given a graphical model $G = (V, E)$ and a family of probability distributions $p \in \mathcal{F}(G, \mathcal{M})$ that factor w.r.t. that distribution. I.e., $f(x) = \sum_{c \in \mathcal{C}} f_c(x_c)$ where $\mathcal{C}$ are a set of cliques.
- Find the closest distribution $p_t$ to $p$ subject to $p_t$ factoring w.r.t. some tree $T = (V, F)$, i.e., $p_t \in \mathcal{F}(T, \mathcal{M})$.
- I.e., optimization problem

$$\begin{aligned} \underset{p_t \in \mathcal{F}(G,\mathcal{M})}{\text{minimize}} \quad & D(p||p_t) \\ \text{subject to} \quad & p_t \in \mathcal{F}(T, \mathcal{M}). \\ & T = (V, F) \text{ is a tree} \end{aligned} \tag{1.11}$$

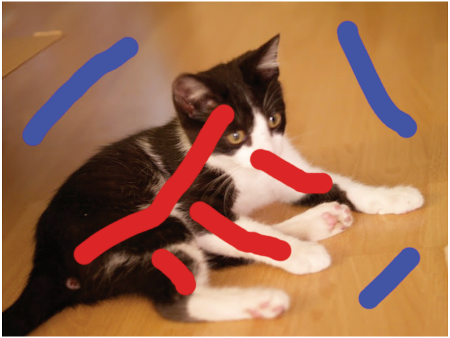
- Discrete problem: Choose the right subset of edges from $E$ that make up a tree (i.e., find a spanning tree of $G$ of best quality).

## Graphical Models: Image Segmentation

- an image needing to be segmented.

# Graphical Models: Image Segmentation

- labeled data in the form of some pixels being marked foreground (red). and others being marked background (blue).



# Graphical Models: Image Segmentation

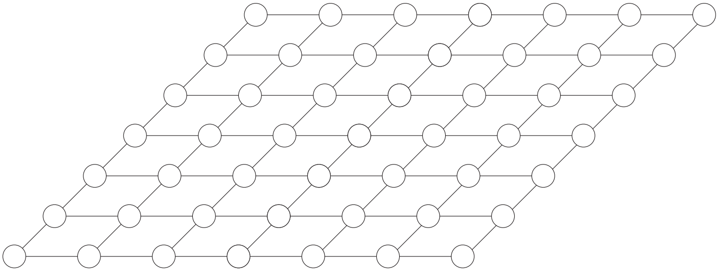
- the foreground is removed from the background.

# Markov random fields and image segmentation

Markov random field

$$\log p(x) \propto \sum_{v \in V(G)} e_v(x_v) + \sum_{(i,j) \in E(G)} e_{ij}(x_i, x_j) \tag{1.12}$$

When $G$ is a 2D grid graph, we have

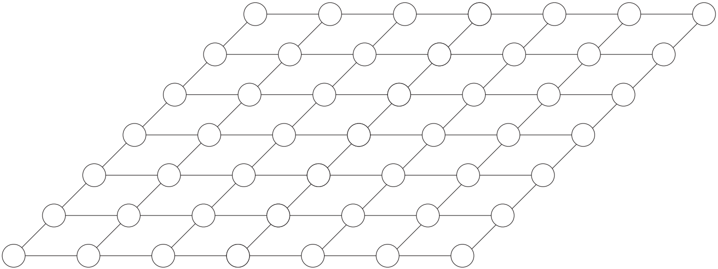


# Markov random fields and image segmentation

- We can create auxiliary graph that involves two new nodes $s$ and $t$ and connect each of $s$ and $t$ to all of the original nodes.
- I.e.,$G_a = (V \cup \{s,t\}, E + \cup_{v \in V}((s,v) \cup (v,t)))$.

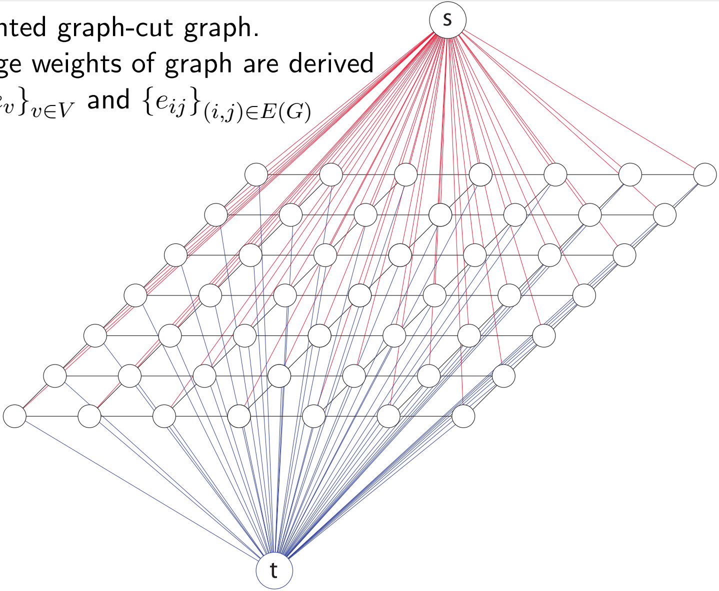
# Markov random fields and image segmentation

Original Graph: $\log p(x) \propto \sum_{v \in V(G)} e_v(x_v) + \sum_{(i,j) \in E(G)} e_{ij}(x_i, x_j)$



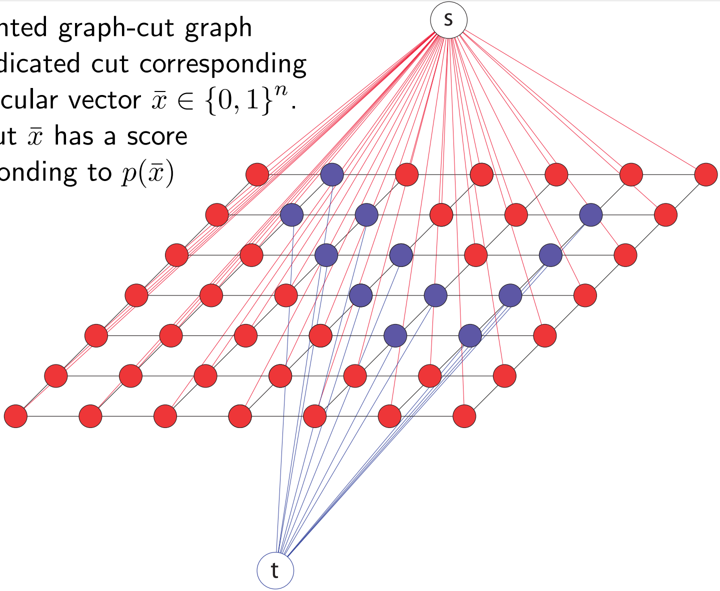
# Markov random fields and image segmentation

Augmented graph-cut graph.
The edge weights of graph are derived from $\{e_v\}_{v\in V}$ and $\{e_{ij}\}_{(i,j)\in E(G)}$



# Markov random fields and image segmentation

Augmented graph-cut graph with indicated cut corresponding to particular vector $\bar{x} \in \{0,1\}^n$. Each cut $\bar{x}$ has a score corresponding to $p(\bar{x})$



# Other applications in or related to computer vision

- Image denoising, total variation, structured convex norms.

$$g(w) = \sum_{i=2}^{N} |w_i - w_{i-1}| \tag{1.13}$$

(from Rodriguez, 2009)

- Multi-label graph cuts
- graphical model inference, computing the Viterbi (or the MPE or the MAP) assignment of a set of random variables.
- Clustering of data sets.

# Diversity Functions

- Diverse web search. Given search term (e.g., "jaguar") but no other information, one probably does not want only articles about cars.

# Diversity Functions

- Diverse web search. Given search term (e.g., "jaguar") but no other information, one probably does not want only articles about cars.
- Given a set $V$ of of items, how do we choose a subset $S \subseteq V$ that is as diverse as possible, with perhaps constraints on $S$ such as its size.

# Diversity Functions

- Diverse web search. Given search term (e.g., “jaguar”) but no other information, one probably does not want only articles about cars.
- Given a set $V$ of of items, how do we choose a subset $S \subseteq V$ that is as diverse as possible, with perhaps constraints on $S$ such as its size.
- How do we choose the smallest set $S$ that maintains a given quality of diversity?

# Diversity Functions

- Diverse web search. Given search term (e.g., "jaguar") but no other information, one probably does not want only articles about cars.
- Given a set $V$ of of items, how do we choose a subset $S \subseteq V$ that is as diverse as possible, with perhaps constraints on $S$ such as its size.
- How do we choose the smallest set $S$ that maintains a given quality of diversity?
- Goal of diversity: ensure proper representation in chosen set that, say otherwise in a random sample, could lead to poor representation of normally underrepresented groups.

# Extractive Document Summarization

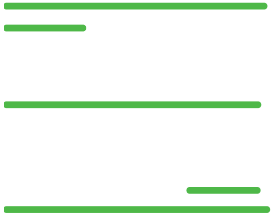
- The figure below represents the sentences of a document

# Extractive Document Summarization

- We extract sentences (green) as a summary of the full document

# Extractive Document Summarization

- We extract sentences (green) as a summary of the full document



# Extractive Document Summarization

- We extract sentences (green) as a summary of the full document

- The summary on the left is a subset of the summary on the right.

# Extractive Document Summarization

- We extract sentences (green) as a summary of the full document


- The summary on the left is a subset of the summary on the right.
- Consider adding a new (blue) sentence to each of the two summaries.

# Extractive Document Summarization

- We extract sentences (green) as a summary of the full document

- The summary on the left is a subset of the summary on the right.
- Consider adding a new (blue) sentence to each of the two summaries.
- The marginal (incremental) benefit of adding the new (blue) sentence to the smaller (left) summary is no more than the marginal benefit of adding the new sentence to the larger (right) summary.

# Extractive Document Summarization

- We extract sentences (green) as a summary of the full document

- The summary on the left is a subset of the summary on the right.
- Consider adding a new (blue) sentence to each of the two summaries.
- The marginal (incremental) benefit of adding the new (blue) sentence to the smaller (left) summary is no more than the marginal benefit of adding the new sentence to the larger (right) summary.
- **diminishing returns** $\leftrightarrow$ **submodularity**

# Web search and information retrieval

- A web search is a form of summarization based on query.
- Goal of a web search engine is to produce a ranked list of web pages that, conditioned on the text query entered, summarizes the most important links on the web.
- Information retrieval (the science of automatically acquiring information), book and music recommendation systems —
- Overall goal: user should quickly find information that is informative, concise, accurate, relevant (to the user's needs), and comprehensive.

# Active Learning and Semi-Supervised Learning

- Given training data $\mathcal{D}_V = \{(x_i, y_i)\}_{i \in V}$ of $(x, y)$ pairs where $x$ is a query (data item) and $y$ is an answer (label), goal is to learn a good mapping $y = h(x)$.

# Active Learning and Semi-Supervised Learning

- Given training data $\mathcal{D}_V = \{(x_i, y_i)\}_{i \in V}$ of $(x, y)$ pairs where $x$ is a query (data item) and $y$ is an answer (label), goal is to learn a good mapping $y = h(x)$.
- Often, getting $y$ is time-consuming, expensive, and error prone (manual transcription, Amazon Turk, etc.)

# Active Learning and Semi-Supervised Learning

- Given training data $\mathcal{D}_V = \{(x_i, y_i)\}_{i \in V}$ of $(x, y)$ pairs where $x$ is a query (data item) and $y$ is an answer (label), goal is to learn a good mapping $y = h(x)$.
- Often, getting $y$ is time-consuming, expensive, and error prone (manual transcription, Amazon Turk, etc.)
- Batch active learning: choose a subset $S \subset V$ so that only the labels $\{y_i\}_{i \in S}$ should be acquired.

# Active Learning and Semi-Supervised Learning

- Given training data $\mathcal{D}_V = \{(x_i, y_i)\}_{i \in V}$ of $(x, y)$ pairs where $x$ is a query (data item) and $y$ is an answer (label), goal is to learn a good mapping $y = h(x)$.
- Often, getting $y$ is time-consuming, expensive, and error prone (manual transcription, Amazon Turk, etc.)
- Batch active learning: choose a subset $S \subset V$ so that only the labels $\{y_i\}_{i \in S}$ should be acquired.
- Adaptive active learning: choose a policy whereby we choose an $i_1 \in V$, get the label $y_{i_1}$, choose another $i_2 \in V$, get label $y_{i_2}$,where each chose can be based on previously acquired labels.

# Active Learning and Semi-Supervised Learning

- Given training data $\mathcal{D}_V = \{(x_i, y_i)\}_{i \in V}$ of $(x, y)$ pairs where $x$ is a query (data item) and $y$ is an answer (label), goal is to learn a good mapping $y = h(x)$.
- Often, getting $y$ is time-consuming, expensive, and error prone (manual transcription, Amazon Turk, etc.)
- Batch active learning: choose a subset $S \subset V$ so that only the labels $\{y_i\}_{i \in S}$ should be acquired.
- Adaptive active learning: choose a policy whereby we choose an $i_1 \in V$, get the label $y_{i_1}$, choose another $i_2 \in V$, get label $y_{i_2}$,where each chose can be based on previously acquired labels.
- Semi-supervised (transductive) learning: Once we have $\{y_i\}_{i \in S}$, infer the remaining labels $\{y_i\}_{i \in V \setminus S}$.

# Markets: Supply Side Economies of scale

- Economies of Scale: Many goods and services can be produced at a much lower per-unit cost only if they are produced in very large quantities.

# Markets: Supply Side Economies of scale

- Economies of Scale: Many goods and services can be produced at a much lower per-unit cost only if they are produced in very large quantities.
- The profit margin for producing a unit of goods improved as more of those goods are created.

# Markets: Supply Side Economies of scale

- Economies of Scale: Many goods and services can be produced at a much lower per-unit cost only if they are produced in very large quantities.
- The profit margin for producing a unit of goods improved as more of those goods are created.
- If you already make a good, making a similar good is easier than if you start from scratch (e.g., Apple making both iPod and iPhone).

# Markets: Supply Side Economies of scale

- Economies of Scale: Many goods and services can be produced at a much lower per-unit cost only if they are produced in very large quantities.
- The profit margin for producing a unit of goods improved as more of those goods are created.
- If you already make a good, making a similar good is easier than if you start from scratch (e.g., Apple making both iPod and iPhone).
- An argument in favor of free trade is that it opens up larger markets to firms in (especially otherwise small markets), thereby enabling better economies of scale, and hence greater efficiency (lower costs and resources per unit of good produced).

# Supply Side Economies of scale: Cost of manufacturing a set of items

- Let $V$ be a set of possible items that a company might possibly wish to manufacture, and let $f(S)$ for $S \subseteq V$ be the cost to that company to manufacture subset $S$.

# Supply Side Economies of scale: Cost of manufacturing a set of items

- Let $V$ be a set of possible items that a company might possibly wish to manufacture, and let $f(S)$ for $S \subseteq V$ be the cost to that company to manufacture subset $S$.
- Ex: $V$ might be colors of paint in a paint manufacturer: green, red, blue, yellow, white, etc.

# Supply Side Economies of scale: Cost of manufacturing a set of items

- Let $V$ be a set of possible items that a company might possibly wish to manufacture, and let $f(S)$ for $S \subseteq V$ be the cost to that company to manufacture subset $S$.
- Ex: $V$ might be colors of paint in a paint manufacturer: green, red, blue, yellow, white, etc.
- Producing green when you are already producing yellow and blue is probably cheaper than if you were only producing some other colors.

$$f(\text{green}, \text{blue}, \text{yellow}) - f(\text{blue}, \text{yellow}) <= f(\text{green}, \text{blue}) - f(\text{blue}) \tag{1.14}$$

# Demand side Economies of Scale: Network Externalities

- consumers of a good derive positive value when size of the market increases.

# Demand side Economies of Scale: Network Externalities

- consumers of a good derive positive value when size of the market increases.
- the value of a network to a user depends on the number of other users in that network.

# Demand side Economies of Scale: Network Externalities

- consumers of a good derive positive value when size of the market increases.
- the value of a network to a user depends on the number of other users in that network.
- Hence, network externalities (Katz & Shapiro 1986) are a form of "demand" economies of scale

# Demand side Economies of Scale: Network Externalities

- consumers of a good derive positive value when size of the market increases.
- the value of a network to a user depends on the number of other users in that network.
- Hence, network externalities (Katz & Shapiro 1986) are a form of "demand" economies of scale
- "value" in this case can be seen as a "willingness-to-pay" for the service (WTP)

## Demand side Economies of Scale: Network Externalities

- consumers of a good derive positive value when size of the market increases.
- the value of a network to a user depends on the number of other users in that network.
- Hence, network externalities (Katz & Shapiro 1986) are a form of "demand" economies of scale
- "value" in this case can be seen as a "willingness-to-pay" for the service (WTP)
- WTP tends to increase but then saturate (like a logistic function)

# Demand side Economies of Scale: Network Externalities

- consumers of a good derive positive value when size of the market increases.
- the value of a network to a user depends on the number of other users in that network.
- Hence, network externalities (Katz & Shapiro 1986) are a form of "demand" economies of scale
- "value" in this case can be seen as a "willingness-to-pay" for the service (WTP)
- WTP tends to increase but then saturate (like a logistic function)
- Given network externalities, a consumer in today's market cares also about the future success of the product and competing products.

# Demand side Economies of Scale: Network Externalities

- consumers of a good derive positive value when size of the market increases.
- the value of a network to a user depends on the number of other users in that network.
- Hence, network externalities (Katz & Shapiro 1986) are a form of "demand" economies of scale
- "value" in this case can be seen as a "willingness-to-pay" for the service (WTP)
- WTP tends to increase but then saturate (like a logistic function)
- Given network externalities, a consumer in today's market cares also about the future success of the product and competing products.
- If the good is durable (or there is human capital investment), the total benefits derived from a good will depend on the number of consumers who adopt compatible products in the future.

# Positive Network Externalities

- railroad - standard rail format and shared access

# Positive Network Externalities

- railroad - standard rail format and shared access
- The telephone, who wants to talk by phone only to oneself?

# Positive Network Externalities

- railroad - standard rail format and shared access
- The telephone, who wants to talk by phone only to oneself?
- the internet, more valuable per person the more people use it.

# Positive Network Externalities

- railroad - standard rail format and shared access
- The telephone, who wants to talk by phone only to oneself?
- the internet, more valuable per person the more people use it.
- ebooks (the more people comment, the better it gets)

# Positive Network Externalities

- railroad - standard rail format and shared access
- The telephone, who wants to talk by phone only to oneself?
- the internet, more valuable per person the more people use it.
- ebooks (the more people comment, the better it gets)
- social network sites: facebook more valuable with everyone online

# Positive Network Externalities

- railroad - standard rail format and shared access
- The telephone, who wants to talk by phone only to oneself?
- the internet, more valuable per person the more people use it.
- ebooks (the more people comment, the better it gets)
- social network sites: facebook more valuable with everyone online
- online education, Coursera, edX, etc. – with many people simultaneously taking a class, all gain due to richer peer discussions due to greater pool of well matched study groups, more simultaneous similar questions/problems that are asked, leading to more efficient learning.

# Positive Network Externalities

- railroad - standard rail format and shared access
- The telephone, who wants to talk by phone only to oneself?
- the internet, more valuable per person the more people use it.
- ebooks (the more people comment, the better it gets)
- social network sites: facebook more valuable with everyone online
- online education, Coursera, edX, etc. – with many people simultaneously taking a class, all gain due to richer peer discussions due to greater pool of well matched study groups, more simultaneous similar questions/problems that are asked, leading to more efficient learning.
- Software, Microsoft office, smartphone apps: more people use it more people report bugs, help with problems, software gets better for every user.

# Positive Network Externalities

- railroad - standard rail format and shared access
- The telephone, who wants to talk by phone only to oneself?
- the internet, more valuable per person the more people use it.
- ebooks (the more people comment, the better it gets)
- social network sites: facebook more valuable with everyone online
- online education, Coursera, edX, etc. – with many people simultaneously taking a class, all gain due to richer peer discussions due to greater pool of well matched study groups, more simultaneous similar questions/problems that are asked, leading to more efficient learning.
- Software, Microsoft office, smartphone apps: more people use it more people report bugs, help with problems, software gets better for every user.
- wikipedia

# Positive Network Externalities

- railroad - standard rail format and shared access
- The telephone, who wants to talk by phone only to oneself?
- the internet, more valuable per person the more people use it.
- ebooks (the more people comment, the better it gets)
- social network sites: facebook more valuable with everyone online
- online education, Coursera, edX, etc. – with many people simultaneously taking a class, all gain due to richer peer discussions due to greater pool of well matched study groups, more simultaneous similar questions/problems that are asked, leading to more efficient learning.
- Software, Microsoft office, smartphone apps: more people use it more people report bugs, help with problems, software gets better for every user.
- wikipedia
- any widely used standard (job training now is useful in the future)

# Positive Network Externalities

- railroad - standard rail format and shared access
- The telephone, who wants to talk by phone only to oneself?
- the internet, more valuable per person the more people use it.
- ebooks (the more people comment, the better it gets)
- social network sites: facebook more valuable with everyone online
- online education, Coursera, edX, etc. – with many people simultaneously taking a class, all gain due to richer peer discussions due to greater pool of well matched study groups, more simultaneous similar questions/problems that are asked, leading to more efficient learning.
- Software, Microsoft office, smartphone apps: more people use it more people report bugs, help with problems, software gets better for every user.
- wikipedia
- any widely used standard (job training now is useful in the future)
- Concepts like the “tipping point”, and “winner take all” markets.

# Other Network Externalities

No Network Externalities

- food/drink - (should be) independent of how many others are eating the type of food.

# Other Network Externalities

No Network Externalities

- food/drink - (should be) independent of how many others are eating the type of food.
- Music - your enjoyment should be independent of others' enjoyment.

# Other Network Externalities

No Network Externalities

- food/drink - (should be) independent of how many others are eating the type of food.
- Music - your enjoyment should be independent of others' enjoyment.

# Other Network Externalities

No Network Externalities

- food/drink - (should be) independent of how many others are eating the type of food.
- Music - your enjoyment should be independent of others' enjoyment.

Negative Network Externalities

- clothing

# Other Network Externalities

No Network Externalities

- food/drink - (should be) independent of how many others are eating the type of food.
- Music - your enjoyment should be independent of others' enjoyment.

Negative Network Externalities

- clothing
- costumes

# Optimization Problem Involving Network Externalities

- (From Mirrokni, Roch, Sundararajan 2012): Let $V$ be a set of users.

# Optimization Problem Involving Network Externalities

- (From Mirrokni, Roch, Sundararajan 2012): Let $V$ be a set of users.
- Let $v_i(S)$ be the value that user $i$ has for a good if $S \subseteq V$ already own the good — e.g. $v_i(S) = \omega_i + f_i(\sum_{j \in S} w_{ij})$ where $\omega_i$ is inherent value, and $f_i$ might be a concave function, and $w_{ij}$is now important $j \in S$ is to $i$ (e.g., a network).

# Optimization Problem Involving Network Externalities

- (From Mirrokni, Roch, Sundararajan 2012): Let $V$ be a set of users.
- Let $v_i(S)$ be the value that user $i$ has for a good if $S \subseteq V$ already own the good — e.g. $v_i(S) = \omega_i + f_i(\sum_{j \in S} w_{ij})$ where $\omega_i$ is inherent value, and $f_i$ might be a concave function, and $w_{ij}$is now important $j \in S$ is to $i$ (e.g., a network).
- Given price $p$ for good, user $i$ buys good if $v_i(S) \geq p$.

# Optimization Problem Involving Network Externalities

- (From Mirrokni, Roch, Sundararajan 2012): Let $V$ be a set of users.
- Let $v_i(S)$ be the value that user $i$ has for a good if $S \subseteq V$ already own the good — e.g. $v_i(S) = \omega_i + f_i(\sum_{j \in S} w_{ij})$ where $\omega_i$ is inherent value, and $f_i$ might be a concave function, and $w_{ij}$is now important $j \in S$ is to $i$ (e.g., a network).
- Given price $p$ for good, user $i$ buys good if $v_i(S) \geq p$.
- We choose initial price $p$ and initial set of users $A \subseteq V$ who get the good for free.

# Optimization Problem Involving Network Externalities

- (From Mirrokni, Roch, Sundararajan 2012): Let $V$ be a set of users.
- Let $v_i(S)$ be the value that user $i$ has for a good if $S \subseteq V$ already own the good — e.g. $v_i(S) = \omega_i + f_i(\sum_{j \in S} w_{ij})$ where $\omega_i$ is inherent value, and $f_i$ might be a concave function, and $w_{ij}$is now important $j \in S$ is to $i$ (e.g., a network).
- Given price $p$ for good, user $i$ buys good if $v_i(S) \geq p$.
- We choose initial price $p$ and initial set of users $A \subseteq V$ who get the good for free.
- Define $S_1 = \{i \notin A : v_i(A) \geq p\}$ initial set of buyers.

# Optimization Problem Involving Network Externalities

- (From Mirrokni, Roch, Sundararajan 2012): Let $V$ be a set of users.
- Let $v_i(S)$ be the value that user $i$ has for a good if $S \subseteq V$ already own the good — e.g. $v_i(S) = \omega_i + f_i(\sum_{j \in S} w_{ij})$ where $\omega_i$ is inherent value, and $f_i$ might be a concave function, and $w_{ij}$is now important $j \in S$ is to $i$ (e.g., a network).
- Given price $p$ for good, user $i$ buys good if $v_i(S) \geq p$.
- We choose initial price $p$ and initial set of users $A \subseteq V$ who get the good for free.
- Define $S_1 = \{i \notin A : v_i(A) \geq p\}$ initial set of buyers.
- $S_2 = \{i \notin A \cup S_1 : v_i(A \cup S_1) \geq p\}$.

# Optimization Problem Involving Network Externalities

- (From Mirrokni, Roch, Sundararajan 2012): Let $V$ be a set of users.
- Let $v_i(S)$ be the value that user $i$ has for a good if $S \subseteq V$ already own the good — e.g. $v_i(S) = \omega_i + f_i(\sum_{j \in S} w_{ij})$ where $\omega_i$ is inherent value, and $f_i$ might be a concave function, and $w_{ij}$is now important $j \in S$ is to $i$ (e.g., a network).
- Given price $p$ for good, user $i$ buys good if $v_i(S) \geq p$.
- We choose initial price $p$ and initial set of users $A \subseteq V$ who get the good for free.
- Define $S_1 = \{i \notin A : v_i(A) \geq p\}$ initial set of buyers.
- $S_2 = \{i \notin A \cup S_1 : v_i(A \cup S_1) \geq p\}$.
- This starts a cascade. Let $S_k = \{\cup_{j<k} S_j \cup A | v_j(\cup_{j<k} S_j \cup A) \geq p\}$,

# Optimization Problem Involving Network Externalities

- (From Mirrokni, Roch, Sundararajan 2012): Let $V$ be a set of users.
- Let $v_i(S)$ be the value that user $i$ has for a good if $S \subseteq V$ already own the good — e.g. $v_i(S) = \omega_i + f_i(\sum_{j \in S} w_{ij})$ where $\omega_i$ is inherent value, and $f_i$ might be a concave function, and $w_{ij}$is now important $j \in S$ is to $i$ (e.g., a network).
- Given price $p$ for good, user $i$ buys good if $v_i(S) \geq p$.
- We choose initial price $p$ and initial set of users $A \subseteq V$ who get the good for free.
- Define $S_1 = \{i \notin A : v_i(A) \geq p\}$ initial set of buyers.
- $S_2 = \{i \notin A \cup S_1 : v_i(A \cup S_1) \geq p\}$.
- This starts a cascade. Let $S_k = \{\cup_{j<k} S_j \cup A | v_j(\cup_{j<k} S_j \cup A) \geq p\}$,
- and let $S_{k^*}$ be the saturation point, lowest value of $k$ such that $S_k = S_{k+1}$

# Optimization Problem Involving Network Externalities

- (From Mirrokni, Roch, Sundararajan 2012): Let $V$ be a set of users.
- Let $v_i(S)$ be the value that user $i$ has for a good if $S \subseteq V$ already own the good — e.g. $v_i(S) = \omega_i + f_i(\sum_{j \in S} w_{ij})$ where $\omega_i$ is inherent value, and $f_i$ might be a concave function, and $w_{ij}$is now important $j \in S$ is to $i$ (e.g., a network).
- Given price $p$ for good, user $i$ buys good if $v_i(S) \geq p$.
- We choose initial price $p$ and initial set of users $A \subseteq V$ who get the good for free.
- Define $S_1 = \{i \notin A : v_i(A) \geq p\}$ initial set of buyers.
- $S_2 = \{i \notin A \cup S_1 : v_i(A \cup S_1) \geq p\}$.
- This starts a cascade. Let $S_k = \{\cup_{j<k} S_j \cup A | v_j(\cup_{j<k} S_j \cup A) \geq p\}$,
- and let $S_{k^*}$ be the saturation point, lowest value of $k$ such that $S_k = S_{k+1}$
- Goal: find $A$ and $p$ to maximize $p \times |S_{k^*}|$.

# Anecdote

From David Brooks, NYTs column, March 28th, 2011 on "Tools for Thinking". In response to Steven Pinker (Harvard) asking a number of people "What scientific concept would improve everybody's cognitive toolkit?"

> *Emergent systems are ones in which many different elements interact. The pattern of interaction then produces a new element that is greater than the sum of the parts, which then exercises a top-down influence on the constituent elements.*

# Submodular Motivation Recap

- Given a set of objects $V = \{v_1, \dots, v_n\}$ and a function $f : 2^V \to \mathbb{R}$ that returns a real value for any subset $S \subseteq V$.
- Suppose we are interested in finding the subset that either maximizes or minimizes the function, e.g., $\operatorname{argmax}_{S \subseteq V} f(S)$, possibly subject to some constraints.
- In general, this problem has exponential time complexity.
- Example: $f$ might correspond to the value (e.g., information gain) of a set of sensor locations in an environment, and we wish to find the best set $S \subseteq V$ of sensors locations given a fixed upper limit on the number of sensors $|S|$.
- In many cases (such as above) $f$ has properties that make its optimization tractable to either exactly or approximately compute.
- One such property is *submodularity*.

# Submodular Definitions

**Definition 1.6.1 (submodular concave)**

A function $f : 2^V \to \mathbb{R}$ is submodular if for any $A, B \subseteq V$, we have that:

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \tag{1.2}$$

An alternate and (as we see in lecture 3) equivalent definition is:

**Definition 1.6.2 (diminishing returns)**

A function $f : 2^V \to \mathbb{R}$ is submodular if for any $A \subseteq B \subset V$, and $v \in V \setminus B$, we have that:

$$f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B) \tag{1.3}$$

This means that the incremental "value", "gain", or "cost" of $v$ decreases (diminishes) as the context in which $v$ is considered grows from $A$ to $B$.

# Subadditive Definitions

**Definition 1.6.1 (subadditive)**

A function $f : 2^V \to \mathbb{R}$ is subadditive if for any $A, B \subseteq V$, we have that:

$$f(A) + f(B) \geq f(A \cup B) \tag{1.15}$$

This means that the "whole" is less than the sum of the parts.

# Supermodular Definitions

**Definition 1.6.2 (supermodular convex)**

A function $f : 2^V \to \mathbb{R}$ is supermodular if for any $A, B \subseteq V$, we have that:

$$f(A) + f(B) \leq f(A \cup B) + f(A \cap B) \tag{1.16}$$

An alternate and equivalent definition is:

**Definition 1.6.3 (increasing returns)**

A function $f : 2^V \to \mathbb{R}$ is supermodular if for any $A \subseteq B \subset V$, and $v \in V \setminus B$, we have that:

$$f(A \cup \{v\}) - f(A) \leq f(B \cup \{v\}) - f(B) \tag{1.17}$$

The incremental "value", "gain", or "cost" of $v$ increases as the context in which $v$ is considered grows from $A$ to $B$.

# Submodular vs. Supermodular

- Submodular and supermodular functions are closely related.

# Submodular vs. Supermodular

- Submodular and supermodular functions are closely related.
- In fact, $f$ is submodular iff $-f$ is supermodular.

# Superadditive Definitions

**Definition 1.6.4 (superadditive)**

A function $f : 2^V \to \mathbb{R}$ is superadditive if for any $A, B \subseteq V$, we have that:

$$f(A) + f(B) \leq f(A \cup B) \tag{1.18}$$

This means that the “whole” is greater than the sum of the parts.

# Modular Definitions

**Definition 1.6.5 (modular)**

A function that is both submodular and supermodular is called **modular**

If $f$ is a modular function, than for any $A, B \subseteq V$, we have

$$f(A) + f(B) = f(A \cap B) + f(A \cup B) \tag{1.19}$$

Modular functions have no interaction, and have value based only on singleton values.

**Proposition 1.6.6**

*If $f$ is modular, it may be written as*

$$f(A) = f(\emptyset) + \sum_{a \in A} \Big( f(\{a\}) - f(\emptyset) \Big) \tag{1.20}$$

# Modular Definitions

## Proof.

We inductively construct the value for $A = \{a_1, a_2, \ldots, a_k\}$.

$$f(a_1) + f(a_2) = f(a_1, a_2) + f(\emptyset) \tag{1.21}$$

$$\text{implies } f(a_1, a_2) = f(a_1) - f(\emptyset) + f(a_2) - f(\emptyset) + f(\emptyset) \tag{1.22}$$

then

$$f(a_1, a_2) + f(a_3) = f(a_1, a_2, a_3) + f(\emptyset) \tag{1.23}$$

$$\text{implies } f(a_1, a_2, a_3) = f(a_1, a_2) - f(\emptyset) + f(a_3) - f(\emptyset) + f(\emptyset) \tag{1.24}$$

$$= f(\emptyset) + \sum_{i=1}^{3} f(a_i) - f(\emptyset) \tag{1.25}$$

□

# Complement function

Given a function $f : 2^V \to \mathbb{R}$, we can find a complement function $\bar{f} : 2^V \to \mathbb{R}$ as $\bar{f}(A) = f(V \setminus A)$ for any $A$.

## Proposition 1.6.7

*$\bar{f}$ is submodular if $f$ is submodular.*

## Proof.

$$\bar{f}(A) + \bar{f}(B) \geq \bar{f}(A \cup B) + \bar{f}(A \cap B) \tag{1.26}$$

follows from

$$f(V \setminus A) + f(V \setminus B) \geq f(V \setminus (A \cup B)) + f(V \setminus (A \cap B)) \tag{1.27}$$

which is true because $V \setminus (A \cup B) = (V \setminus A) \cap (V \setminus B)$ and $V \setminus (A \cap B) = (V \setminus A) \cup (V \setminus B)$. $\square$

# Submodularity

- Submodular functions have a long history in economics, game theory, combinatorial optimization, electrical networks, and operations research.
- They are gaining importance in machine learning as well (one of our main motivations for offering this course).
- Arbitrary set functions are hopelessly difficult to optimize, while the minimum of submodular functions can be found in polynomial time, and the maximum can be constant-factor approximated in low-order polynomial time.
- Submodular functions share properties in common with both convex and concave functions.

# Attractions of Convex Functions

Why do we like Convex Functions? (Quoting Lovász 1983):

1. *Convex functions occur in many mathematical models in economy, engineering, and other sciences. Convexity is a very natural property of various functions and domains occurring in such models; quite often the only non-trivial property which can be stated in general.*

# Attractions of Convex Functions

Why do we like Convex Functions? (Quoting Lovász 1983):

1. *Convex functions occur in many mathematical models in economy, engineering, and other sciences. Convexity is a very natural property of various functions and domains occurring in such models; quite often the only non-trivial property which can be stated in general.*
2. *Convexity is preserved under many natural operations and transformations, and thereby the effective range of results can be extended, elegant proof techniques can be developed as well as unforeseen applications of certain results can be given.*

# Attractions of Convex Functions

Why do we like Convex Functions? (Quoting Lovász 1983):

1. *Convex functions occur in many mathematical models in economy, engineering, and other sciences. Convexity is a very natural property of various functions and domains occurring in such models; quite often the only non-trivial property which can be stated in general.*
2. *Convexity is preserved under many natural operations and transformations, and thereby the effective range of results can be extended, elegant proof techniques can be developed as well as unforeseen applications of certain results can be given.*
3. *Convex functions and domains exhibit sufficient structure so that a mathematically beautiful and practically useful theory can be developed.*

# Attractions of Convex Functions

Why do we like Convex Functions? (Quoting Lovász 1983):

1. *Convex functions occur in many mathematical models in economy, engineering, and other sciences. Convexity is a very natural property of various functions and domains occurring in such models; quite often the only non-trivial property which can be stated in general.*
2. *Convexity is preserved under many natural operations and transformations, and thereby the effective range of results can be extended, elegant proof techniques can be developed as well as unforeseen applications of certain results can be given.*
3. *Convex functions and domains exhibit sufficient structure so that a mathematically beautiful and practically useful theory can be developed.*
4. *There are theoretically and practically (reasonably) efficient methods to find the minimum of a convex function.*

# Attractions of Submodular Functions

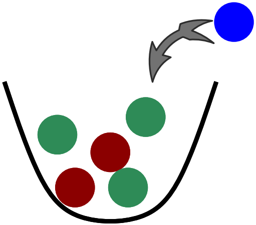
In this course, we wish to demonstrate that submodular functions also possess attractions of these four sorts as well.

# Example Submodular: Number of Colors of Balls in Urns

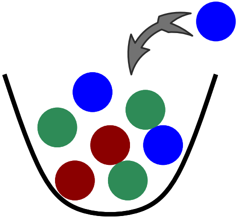
- Consider an urn containing colored balls. Given a set $S$ of balls, $f(S)$ counts the number of distinct colors.

# Example Submodular: Number of Colors of Balls in Urns

- Consider an urn containing colored balls. Given a set $S$ of balls, $f(S)$ counts the number of distinct colors.



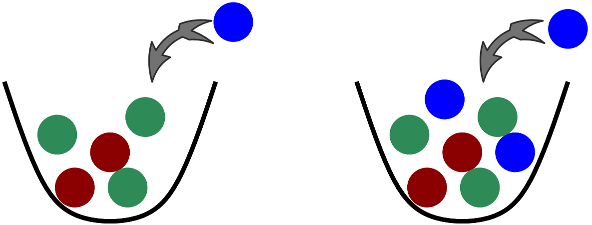
Initial value: 2 (colors in urn).
New value with added blue ball: 3



Initial value: 3 (colors in urn).
New value with added blue ball: 3

# Example Submodular: Number of Colors of Balls in Urns

- Consider an urn containing colored balls. Given a set $S$ of balls, $f(S)$ counts the number of distinct colors.



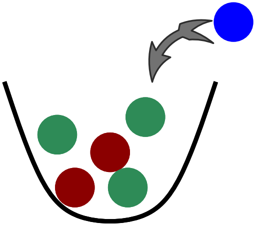
Initial value: 2 (colors in urn).
New value with added blue ball: 3

Initial value: 3 (colors in urn).
New value with added blue ball: 3

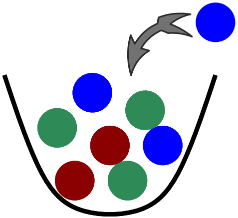
- Submodularity: Incremental Value of Object Diminishes in a Larger Context (diminishing returns).

# Example Submodular: Number of Colors of Balls in Urns

- Consider an urn containing colored balls. Given a set $S$ of balls, $f(S)$ counts the number of distinct colors.



Initial value: 2 (colors in urn).
New value with added blue ball: 3



Initial value: 3 (colors in urn).
New value with added blue ball: 3

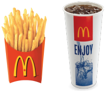
- Submodularity: Incremental Value of Object Diminishes in a Larger Context (diminishing returns).
- Thus, $f$ is submodular.

# Ex. Submodular: Consumer Costs of Living

- Consumer costs are very often submodular.

# Ex. Submodular: Consumer Costs of Living

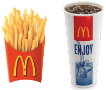
- Consumer costs are very often submodular. For example:

$f($ 

$) + f($  $) \geq f($  $) + f($  $)$

# Ex. Submodular: Consumer Costs of Living

- Consumer costs are very often submodular. For example:

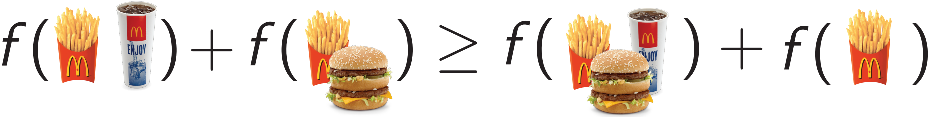
$f($ 

$) + f($  $) \geq f($  $) + f($  $)$

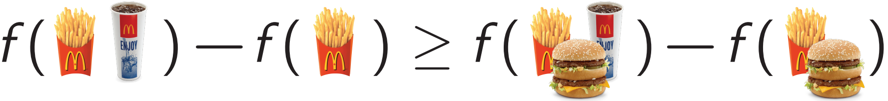
- Rearranging terms, we can see this as diminishing returns:

# Ex. Submodular: Consumer Costs of Living

- Consumer costs are very often submodular. For example:

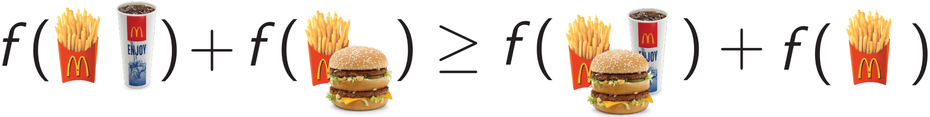


- Rearranging terms, we can see this as diminishing returns:

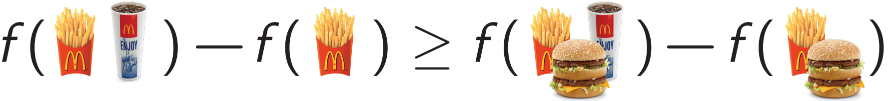


# Ex. Submodular: Consumer Costs of Living

- Consumer costs are very often submodular. For example:




- Rearranging terms, we can see this as diminishing returns:

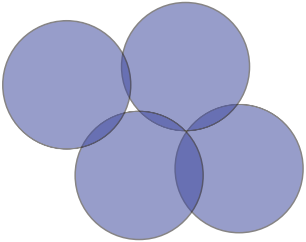



- This is very common: The additional cost of a coke is, say, free if you add it to fries and a hamburger, but when added just to an order of fries, the coke is not free.

# Area of the union of areas indexed by $A$

- Let $V$ be a set of indices, and each $v \in V$ indexes a given sub-area of some region. Let area$(v)$ be the area corresponding to item $v$.
- Let $f(S) = \bigcup_{s \in S}$ area$(s)$ be the union of the areas indexed by elements in $A$.
- Then $f(S)$ is submodular.

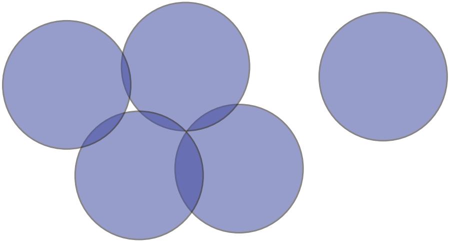
# Area of the union of areas indexed by $A$



Union of areas of elements of $A$ is given by:

$$f(A) = f(\{a_1, a_2, a_3, a_4\})$$

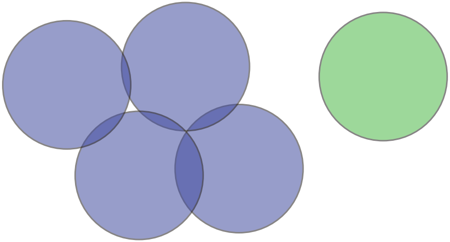
# Area of the union of areas indexed by $A$



Area of $A$ along with with $v$:

$$f(A \cup \{v\}) = f(\{a_1, a_2, a_3, a_4\} \cup \{v\})$$

# Area of the union of areas indexed by $A$

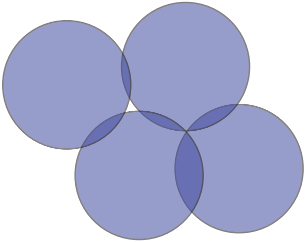


Gain (value) of $v$ in context of $A$:

$$f(A \cup \{v\}) - f(A) = f(\{v\})$$

We get full value $f(\{v\})$ in this case since the area of $v$ has no overlap with that of $A$.

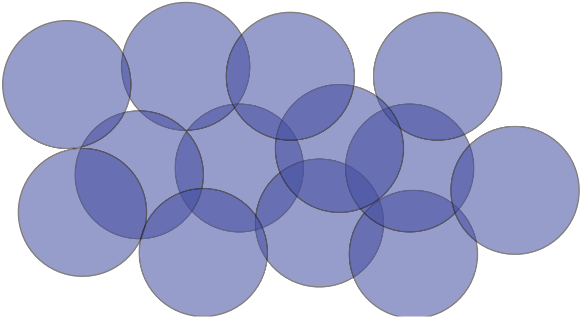
# Area of the union of areas indexed by $A$



Area of $A$ once again.

$$f(A) = f(\{a_1, a_2, a_3, a_4\})$$

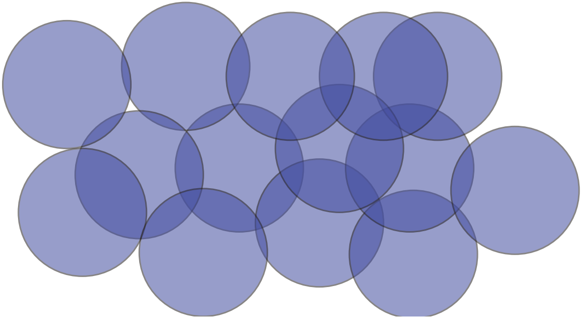
# Area of the union of areas indexed by $A$



Union of areas of elements of $B \supset A$, where $v$ is not included:

$$f(B) \text{ where } v \notin B \text{ and where } A \subseteq B$$

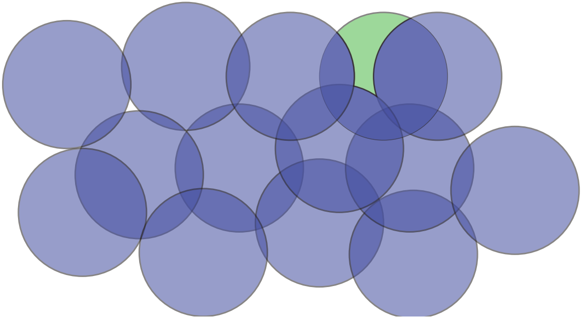
# Area of the union of areas indexed by $A$



Area of $B$ now also including $v$:

$$f(B \cup \{v\})$$

# Area of the union of areas indexed by $A$



Incremental value of $v$ in the context of $B \supset A$.

$$f(B \cup \{v\}) - f(B) < f(\{v\}) = f(A \cup \{v\}) - f(A)$$

So benefit of $v$ in the context of $A$ is greater than the benefit of $v$ in the context of $B \supseteq A$.

# Example Submodular: Entropy from Information Theory

- Entropy is submodular. Let $V$ be the index set of a set of random variables, then the function

$$f(A) = H(X_A) = -\sum_{x_A} p(x_A) \log p(x_A) \tag{1.28}$$

is submodular.

- Proof: conditioning reduces entropy. With $A \subseteq B$ and $v \notin B$,

$$H(X_v|X_B) = H(X_{B+v}) - H(X_B) \tag{1.29}$$
$$\leq H(X_{A+v}) - H(X_A) = H(X_v|X_A) \tag{1.30}$$

# Example Submodular: Entropy from Information Theory

- Alternate Proof: Conditional mutual Information is always non-negative.
- Given $A, B, C \subseteq V$, consider conditional mutual information quantity:

$$I(X_{A\setminus B}; X_{B\setminus A}|X_{A\cap B}) = \sum_{x_{A\cup B}} p(x_{A\cup B}) \log \frac{p(x_{A\setminus B}, x_{B\setminus A}|x_{A\cap B})}{p(x_{A\setminus B}|x_{A\cap B})p(x_{B\setminus A}|x_{A\cap B})}$$
$$= \sum_{x_{A\cup B}} p(x_{A\cup B}) \log \frac{p(x_{A\cup B})p(x_{A\cap B})}{p(x_A)p(x_B)} \geq 0 \tag{1.31}$$

then

$$I(X_{A\setminus B}; X_{B\setminus A}|X_{A\cap B})$$
$$= H(X_A) + H(X_B) - H(X_{A\cup B}) - H(X_{A\cap B}) \geq 0 \tag{1.32}$$

so entropy satisfies

$$H(X_A) + H(X_B) \geq H(X_{A\cup B}) + H(X_{A\cap B}) \tag{1.33}$$

# Example Submodular: Mutual Information

- Also, symmetric mutual information is submodular,

$$f(A) = I(X_A; X_{V \setminus A}) = H(X_A) + H(X_{V \setminus A}) - H(X_V) \quad (1.34)$$

Note that $f(A) = H(X_A)$ and $\bar{f}(A) = H(X_{V \setminus A})$, and adding submodular functions preserves submodularity (which we will see quite soon).