Data Mining and Inormation Diffusion

DM2 – Ricerca degli Innovatori

Material integrated by Jure Leskovec, Kleinberg book: Networks, Crowds and markets, tesi di laurea Walter Tocci

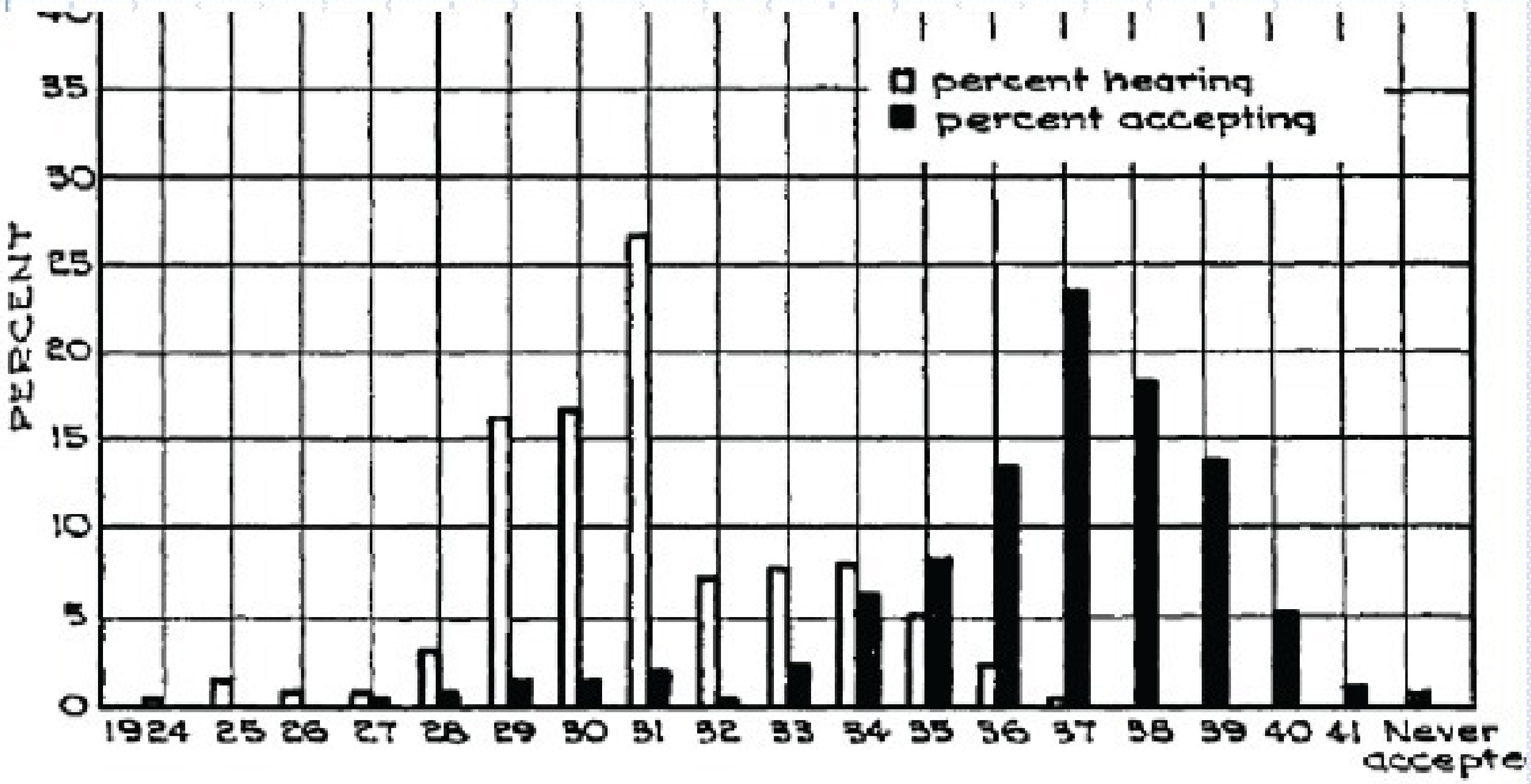
Information diffusion

- Anything that propagates over a network comes under the umbrella of "information diffusion."

- A fundamental process in social networks:
  Behaviors that cascade from node to node like an epidemic
  - News, opinions, rumors, fads, urban legends, …
  - Word-of-mouth effects in marketing: rise of new websites, free web based services
  - Virus, disease propagation
  - Change in social priorities: smoking, recycling
  - Saturation news coverage: topic diffusion among bloggers
  - Internet-energized political campaigns
  - Cascading failures in financial markets
  - Localized effects: riots, people walking out of a lecture

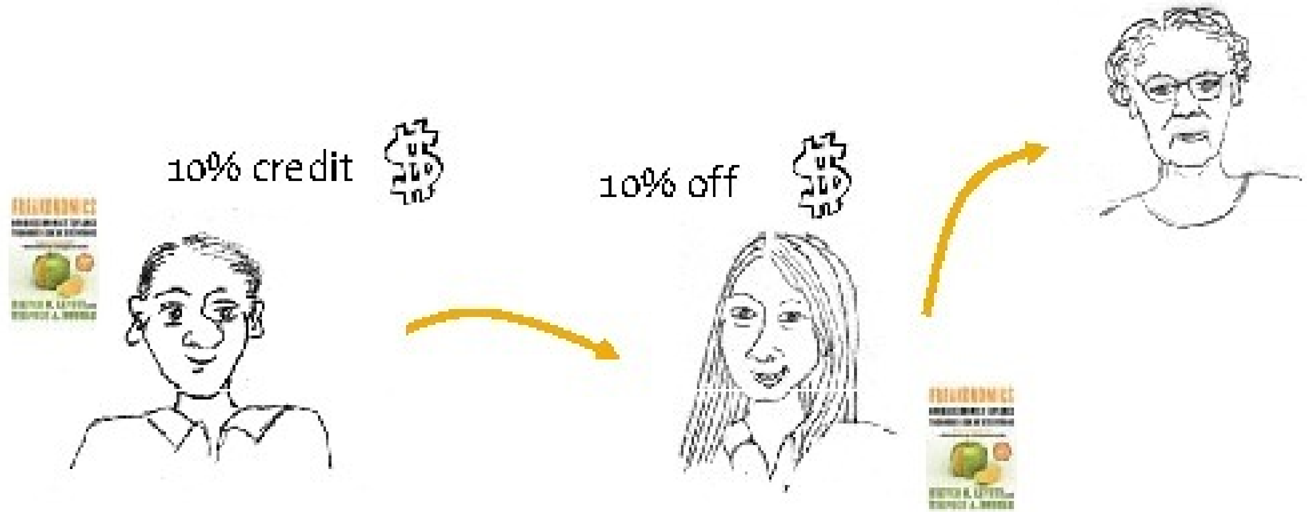
- Experimental studies of diffusion:
  - Spread of new agricultural practices [Ryan-Gross 1943]
    - Adoption of a new hybrid-corn between the 259 farmers in Iowa
    - Classical study of diffusion
    - Interpersonal network plays important role in adoption
      → Diffusion is a social process
  - Spread of new medical practices [Coleman et al. 1966]
    - Studied the adoption of a new drug between doctors in Illinois
    - Clinical studies and scientific evaluations were not sufficient to convince the doctors
    - It was the social power of peers that led to adoption

- Spreading through networks:
  - Cascading behavior
  - Diffusion of innovations
  - Epidemics
- Examples:
  - Biological:
    - Diseases via contagion
  - Technological:
    - Cascading failures
    - Spread of information
  - Social:
    - Rumors, news, new technology
    - Viral marketing

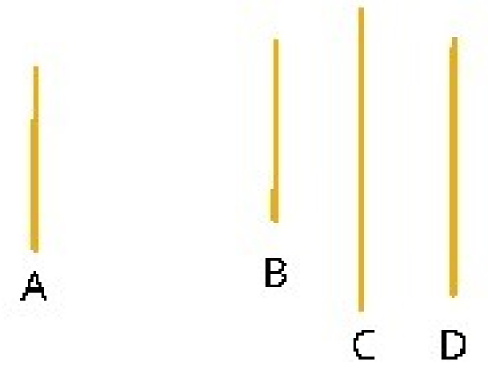


Diffusion is a social process

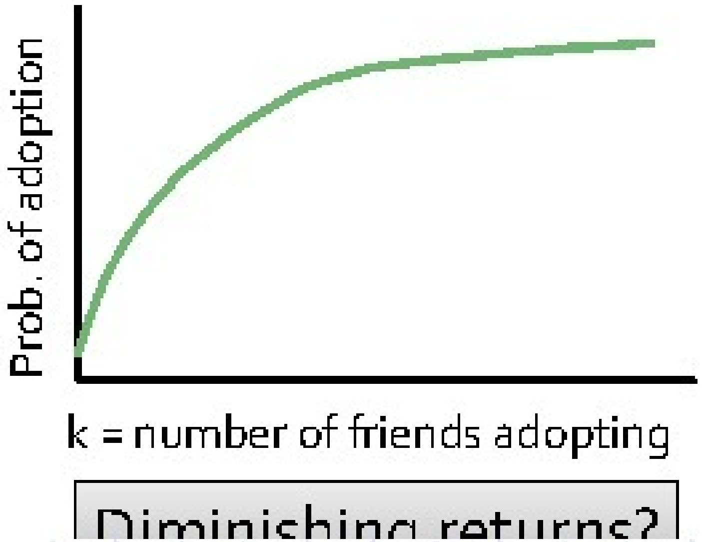
- Senders and followers of recommendations receive discounts on products

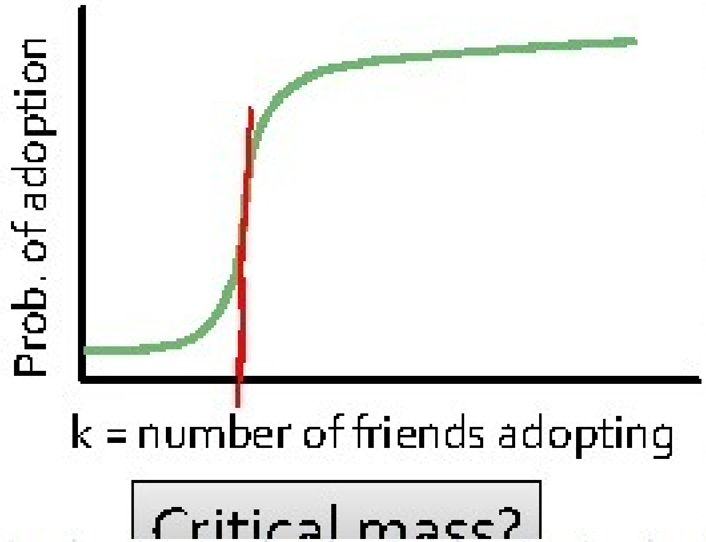


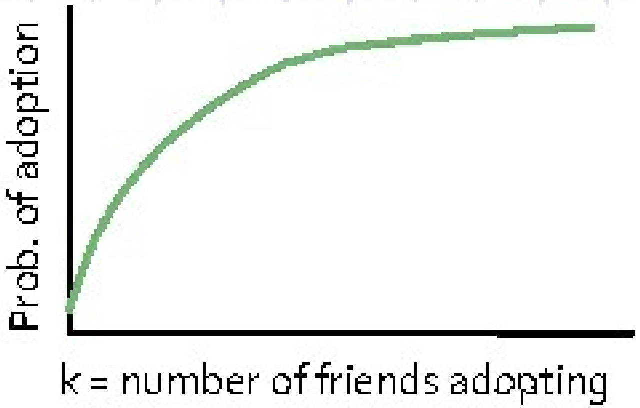
- Diffusion has many (very interesting) flavors:
  - The contagion of obesity [Christakis et al. 2007]
    - If you have an overweight friend your chances of becoming obese increases by 57%
  - Psychological effects of others' opinions, *e.g.*: Which line is closest in length to A? [Asch 1958]



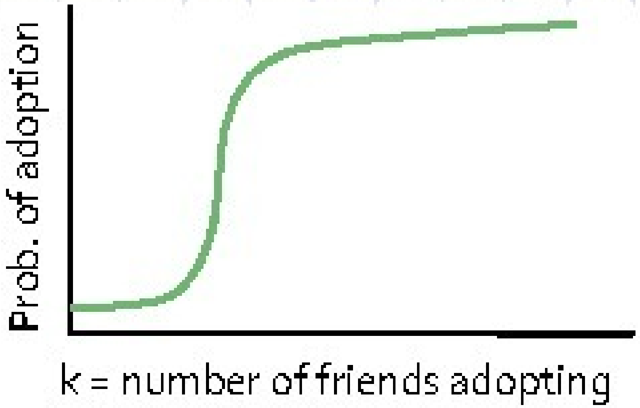
- Basis for models:
  - Probability of adopting new behavior depends on the number of friends who have adopted [Bass ‘69, Granovetter ‘78, Shelling ’78]
- What’s the dependence?







Diminishing returns?



Critical mass?

- Key issue: qualitative shape of diffusion curves
  - Diminishing returns? Critical mass?
  - Distinction has consequences for models of diffusion at population level

- Probabilistic models:
  - Example:
    - "catch" a disease with some prob. from neighbors in the network
- Decision based models:
  - Example:
    - Adopt new behaviors if *k* of your friends do

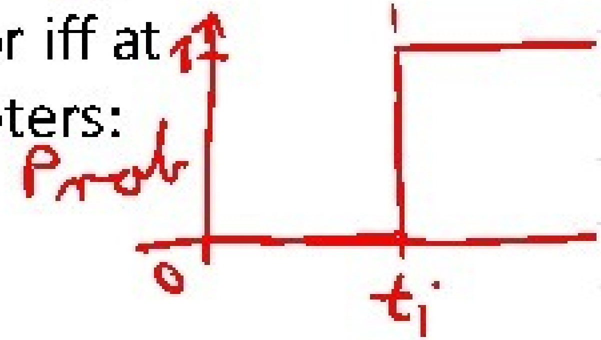
- Two flavors, two types of questions:
  - A) Probabilistic models: Virus Propagation
    - SIS: Susceptible–Infective–Susceptible (*e.g.*, flu)
    - SIR: Susceptible–Infective–Recovered (*e.g.*, chicken-pox)
    - Question: Will the virus take over the network?
    - Independent contagion model
  - B) Decision based models: Diffusion of Innovation
    - Threshold model
    - Herding behavior
    - Questions:
      - Finding influential nodes
      - Detecting cascades

- Influence of actions of others
  - Model where everyone sees everyone else's behavior
- Sequential decision making
  - Picking a restaurant:
    - Consider you are choosing a restaurant in an unfamiliar town
    - Based on Yelp reviews you intend to go to restaurant A
    - But then you arrive there is no one eating at A but the next door restaurant B is nearly full
  - What will you do?
    - Information that you can infer from other's choices may be more powerful than your own

- Herding:
  - There is a decision to be made
  - People make the decision sequentially
  - Each person has some private information that helps guide the decision
  - You can’t directly observe private info of others but can see what they do
    - Can make inferences about their private information

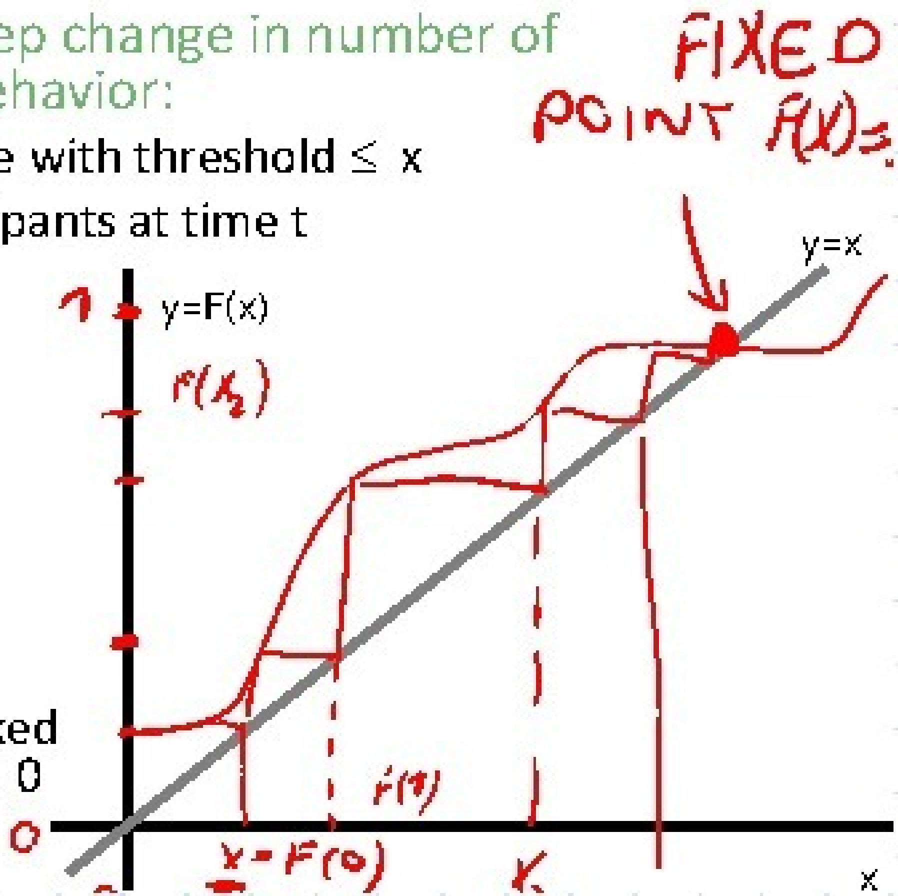
- Collective action [Granovetter, '78]
  - Model where everyone sees everyone else's behavior
  - Examples:
    - Clapping or getting up and leaving in a theater
    - Keeping your money or not in a stock market
    - Neighborhoods in cities changing ethnic composition
    - Riots, protests, strikes

- n people – everyone observes all actions
- Each person *i* has a threshold t_i
  - Node *i* will adopt the behavior iff at least t_i other people are adopters:
    - Small t_i: early adopter
    - Large t_i: late adopter



- The population is described by $\{t_1, \ldots, t_n\}$
  - F(x) … fraction of people with threshold $t_i \leq x$

- Think of the step-by-step change in number of people adopting the behavior:
  - F(x) … fraction of people with threshold ≤ x
  - s(t) … number of participants at time t
- Easy to simulate:
  - s(0) = 0
  - s(1) = F(0)
  - s(2) = F(s(1)) = F(F(0))
  - s(t+1) = F(s(t)) = $F^{t+1}(0)$
- F(x)=x – stable point
  - There could be other fixed points but starting from 0 we never reach them



- It does not take into account:
  - No notion of social network – more influential users
  - It matters who the early adopters are, not just how many
  - Models people's awareness of size of participation not just actual number of people participating
  - Modeling thresholds
    - Richer distributions
    - Deriving thresholds from mode basic assumptions
      - game theoretic models

It does not take into account:

- Modeling perceptions of who is adopting the behavior/ who you believe is adopting
- Non monotone behavior – dropping out if too many people adopt
- Similarity – thresholds not based only on numbers
- People get "*locked in*" *to* certain choice over a period of time

Diffusion of innovation

- is a theory that seeks to explain how, why, and at what rate new ideas and technology spread through cultures.
- Everett Rogers, a professor of rural sociology, popularized the theory in his 1962 book Diffusion of Innovations.
- He said diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social system.
- The origins of the diffusion of innovations theory are varied and span multiple disciplines.

Chi sono gli innovatori

Innovatori

- Innovators
- Early Adopter
- Early Majority
- Late Majority
- Laggard

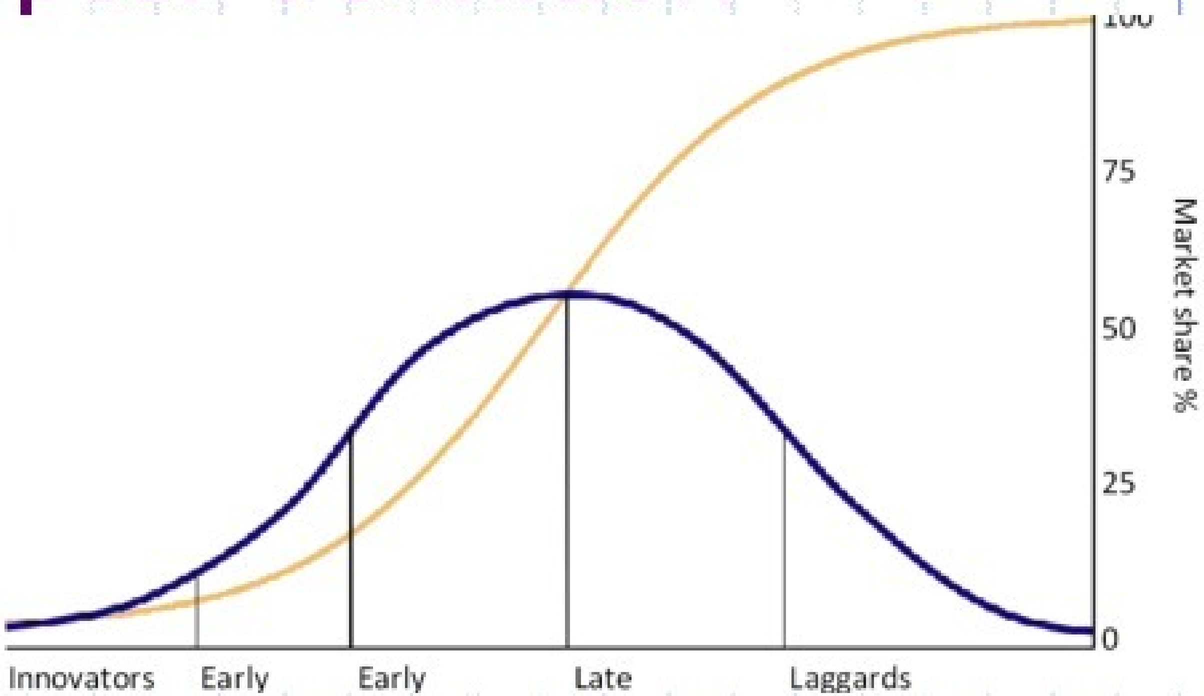
- Istruzione, propensione al rischio, informazione, velocit`a del processo di decision making

Importanza degli inovatori

- Anticipatori
- Precursori
- Aiutano a conferire l'immagine di opinion leader
- Permettono una correzione delle caratteristiche del prodotto
- on-time Heavy-user
- ⇒ sono importanti per una efficace campagna di marketing e di customer satisfaction

- Problematiche e questioni irrisolte appiattimento degli strumenti utilizzati (modelli
- matematico-statistici) analisi su campioni di piccole dimensioni ricerca degli Innovatori come categoria, non come individui risutati spesso contrastanti, non raggiungono un accordo solo primi acquisti

Roger Adopter definition



- $\bar{x} - 2\sigma$, tra Innovatori ed Early Adopter;
- $\bar{x} - \sigma$, tra Early Adopter ed Early Majority;
- $\bar{x}$, tra Early Majority e Late Majority;
- $\bar{x} + \sigma$, tra Late Majority e Laggard.

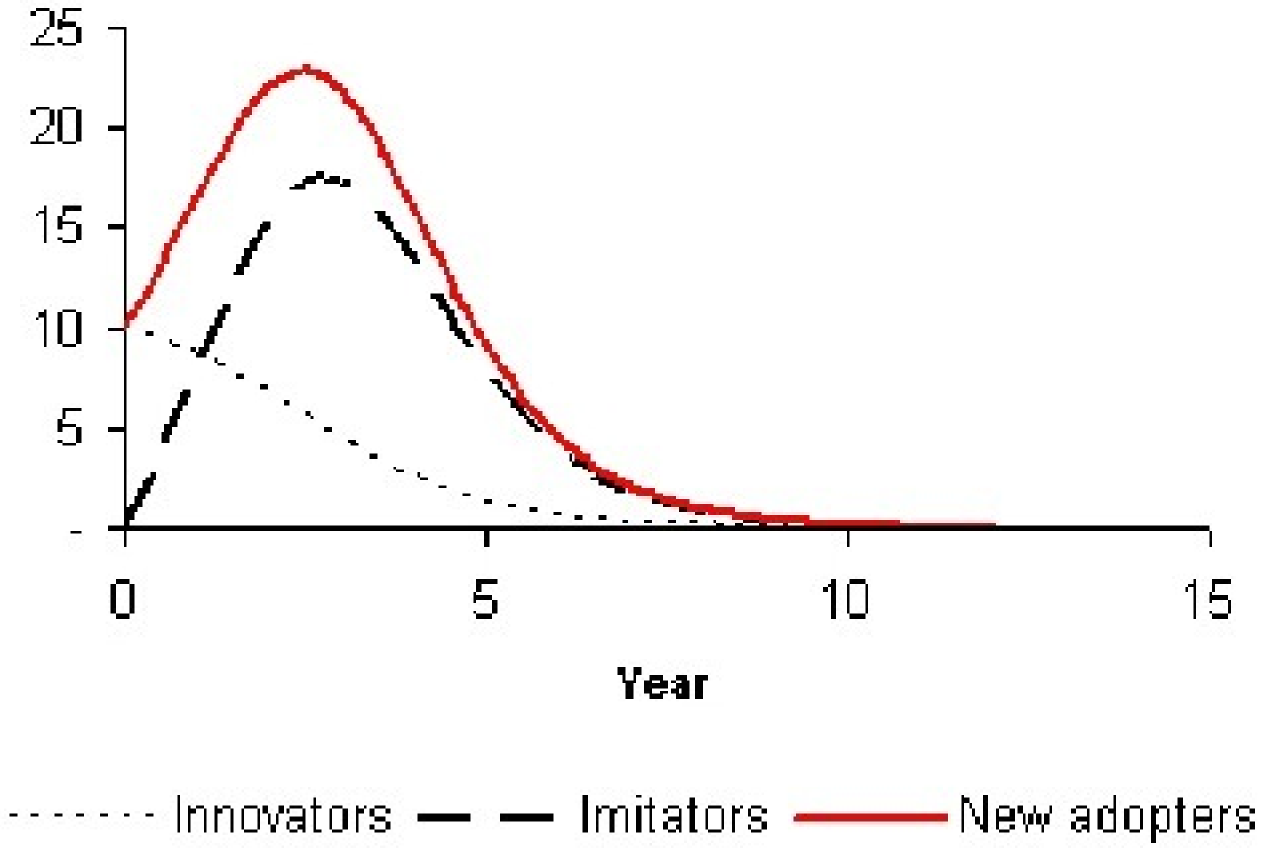
Bass Diffusion Model

- le potenzialità di mercato, vale a dire il numero totale di persone che possono adottare l'innovazione;
- il coefficiente di influenza esterna (o di innovazione), vale a dire la probabilità che qualcuno che ancora non sta adottando l'innovazione inizi a farlo sotto l'influenza dei mass-media o di altri fattori esterni;
- il coefficiente di influenza interna (o di imitazione), che racchiude la probabilità che qualcuno che ancora non sta adottando l'innovazione inizi a farlo sulla base del passa-parola o di altre forme di influenza diretta da parte di chi sta già utilizzando il prodotto.

- Rogers - definizione Adopters
- Bass - Bass Diffusion Model P (t) = p + (q/m) Y (t)
- Dove:

dove p e q/m sono costanti, m è il potenziale di mercato e $Y(t)$ è il numero di acquirenti precedenti al tempo t. p e q sono i *coefficienti di innovazione ed imitazione.*

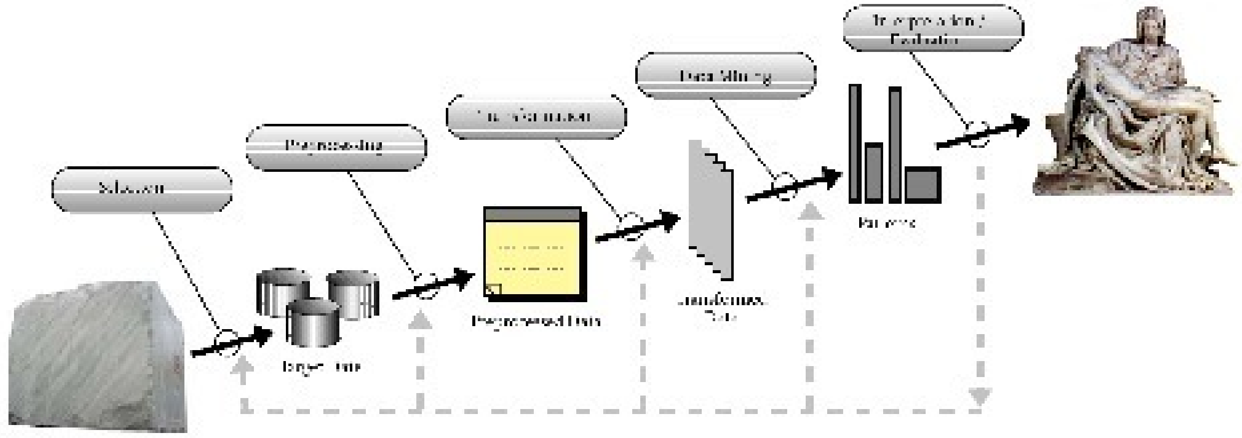
Bass Model



Nuova definizione di adopter

- Gli innovatori sono coloro che permettono di prevedere in anticipo **l’andamento tipico** di un prodotto:
- Come trovare l’andamento tipico di un prodotto
- L’innovatore è colui/lei che propende ad adottare in anticipo l’andamento di un prodotto

Il processo di Knowledge Discovery and Data Mining



- Approccio tipo KDD
- SPM come particolare tecnica di DM

utente	settimana 1	settimana 2	settimana 3	
A	acquisto 1	acquisto 5	acquisto 3	←
B	acquisto 1	acquisto 3	acquisto 3	
C	acquisto 1	acquisto 5	acquisto 3	←

Tabella: Sequenze in un time-series Database

- Prodotto Nuovo
- Sequenza Tipica
- Adopter → Adopter Tipico
- Innovatori → Innovatore Tipico
- TOC - Tasso di Omogeneità del Comportamento

$$TOC_j(i) = \frac{num_j(i)}{num_{Adopter}(i)}$$

$$i = i - esimo \text{ individuo}$$
$$j = \text{categoria di Adopter (Tipico)}$$

La Ricerca degli Innovatori Tipici

- Problema di Ricerca degli Innovatori Tipici (**RIT**)
- Metodologia di risoluzione al problema RIT:
  1. Sequential Pattern Mining → Ricerca delle **Sequenze Tipiche**
  2. Ricerca degli **Innovatori Tipici**
  3. **Calcolo del TOC** per diverse categorie di prodotti e clienti
  4. Ricerca del **TOC massimo**

Fase del processo KDD in pratica

- Selezione
- Preprocessing
- Trasformazione
- Data Mining

- Selezione Prodotti Nuovi
- Calcolo delle quantità prodotti a peso
- Discretizzazione del tempo
- Scelta degli *items*

Trasformazione: scelta degli items

- Quantità?

$$A : < \ 1, \ 2, \ 2, \ 1 \ >$$

$$B : < \ 2, \ 5, \ 5, \ 4 \ >$$

- Andamenti?

$$< \ 1, \ 0, \ -1 \ >$$

- 1: andamento **crescente**
- 0: andamento **costante**
- -1: andamento **decrescente**

$$< \ crescente, \ costante, \ decrescente \ >$$

- Selezione
- Preprocessing
- Trasformazione
- Data Mining

- Macro-analisi (regressione)
- Micro-analisi (Sequential Pattern Mining)

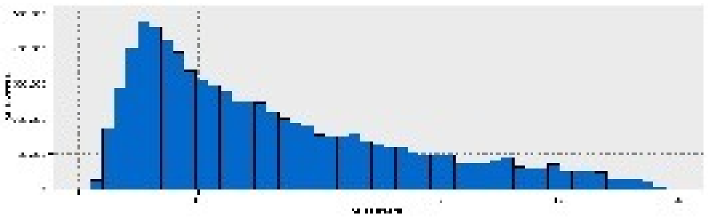


Figura: Andamento Acquisti Ripetuti

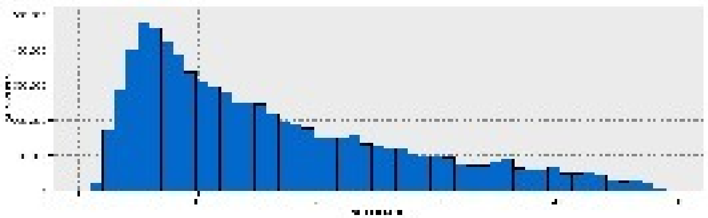
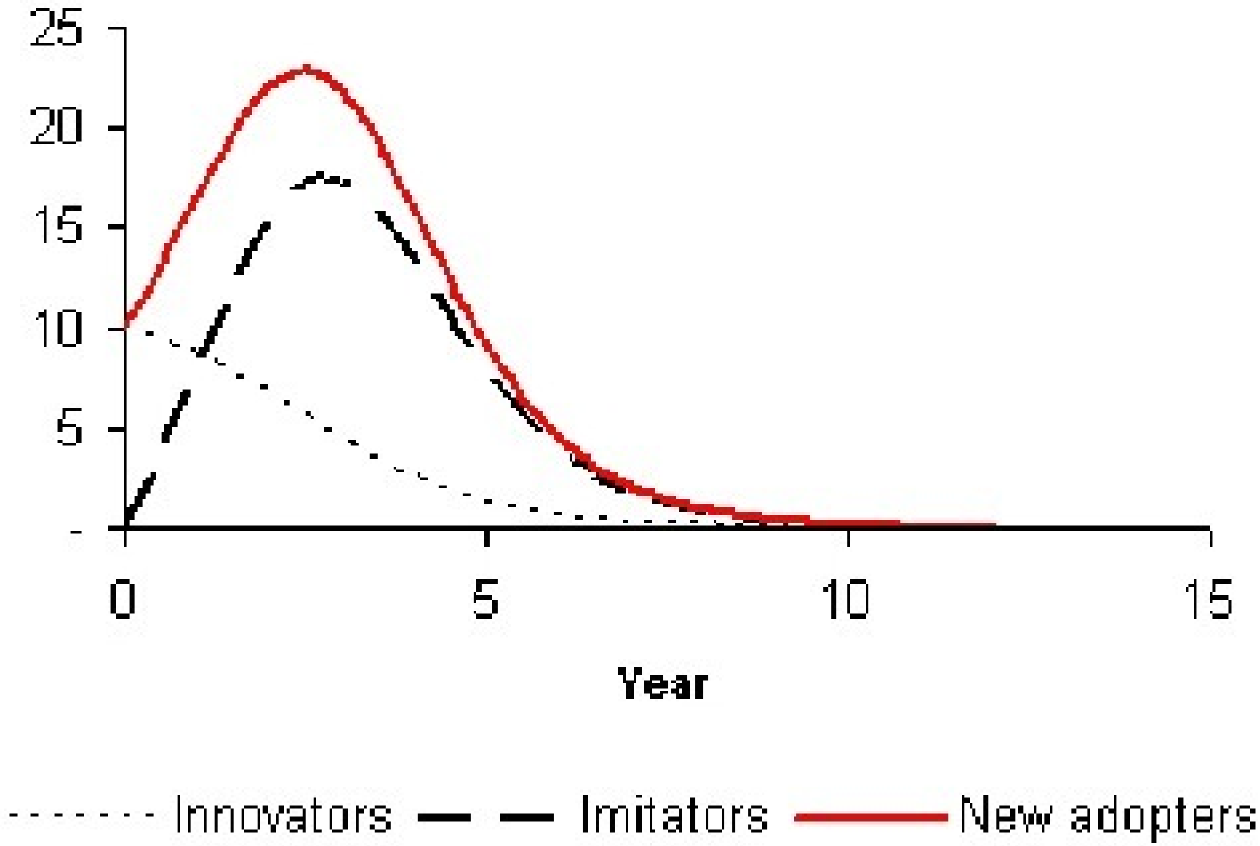


Figura: Acquisti Prodotti Nuovi allineati: Febbraio-Settembre

Bass Model



Primi Aquisti Vs Acquisti Ripetuti - in dettaglio

Grado	b	c	d	e	f	g	h	Correlaz.
1	-4.6658	-	-	-	-	-	-	0.2576
2	4.2681	-2.6304	-	-	-	-	-	0.6835
3	1.6032	-1.9727	6.3322	-	-	-	-	0.9409
4	2.9348	-4.8752	3.0918	6.8294	-	-	-	0.9847
5	2.2280	-2.4282	-3.6792	1.4449	-4.7285	-	-	0.9881
6	2.8639	6.6243	-1.8925	2.0069	-9.4523	1.6628	-	0.9948
7	-1.1272	1.4937	-4.1635	5.2610	-3.4664	1.1651	-1.5855	0.9957

Tabella: Coefficienti di regressione Primi Acquisti

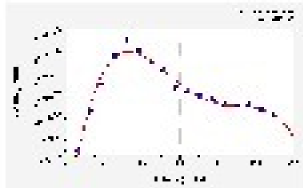


Figura: Regressione grado 4 su Primi Acquisti

Grado	b	c	d	e	f	g	h	Correlaz.
1	-2.6119	-	-	-	-	-	-	0.1152
2	6.6306	-3.6288	-	-	-	-	-	0.7450
3	2.1448	-2.3833	7.4090	-	-	-	-	0.9593
4	3.3407	-5.1790	3.1090	-6.5780	-	-	-	0.9850
5	2.2536	-1.4152	-2.2123	2.6150	-7.2729	-	-	0.9901
6	-7.5493	9.4437	-2.4473	2.4955	-1.1498	1.9947	-	0.9963
7	-1.2760	1.6504	-4.3760	5.2591	-3.2910	1.0478	-1.3465	0.9967

Tabella: Coefficienti di regressione Acquisti Ripetuti

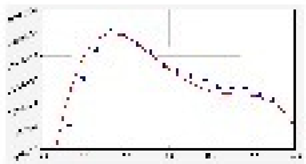


Figura: Predizione su Ottobre

delle Sequenze Tipiche

Prodotto	Sequenza Tipica	Supporto
Paste Fresche	1 0 0	32.9%
Pementa	1 0 0	27.4%
Piadina all'olio	0 0 0	67.7%
Bevanda alla Soya	1 0 0	36.3%
Yogurt	0 0 0	68.2%
Ricotta	1 0 0	41.2%
Succo di Frutta	0 0 0	69.4%
Birra	0 0 0	42.9%
Salame	1 0 0	38.3%
Base Pizza	1 -1 1	29.2%
Peperoncino	1 0 0	28.1%
Tacchino Arrosto	1 0 0	36.8%
Cocomero a cubetti	1 1 -1	31.2%
Insalata di Riso	1 1 -1	35.3%
Prosciutto Spalmabile	0 0 0	66.1%
Susine President	1 0 0	27.4%
Latte UHT	1 0 0	37.7%
Bevanda al Limone	0 0 0	51.3%
Prosciutto Cotto Alta Qualità	1 0 0	42.6%
Bocconcino	1 -1 1	40.1%
Aglio Biologico	1 0 0	82.7%
Limoni	0 0 0	37.9%

Prodotto	Sequenza Tipica	Supporto
Pane Bianco Confezionato	0 0 0	83.3%
Nettarine Gialle	1 0 0	37.4%
Fragole	1 0 0	40.2%
Patate Novelle	1 0 1	36.4%
Fagiolini Verdi	0 0 0	40.2%
Salmone Affumicato	1 0 0	54.4%
Prosciutto Crudo	1 0 0	43.3%
Verdure Precotte	1 0 0	31.7%
Vitellone	1 0 0	37.0%
Dessert al Limone	1 0 0	41.6%
Vino	0 0 0	67.5%
Insetticida	1 0 0	35.5%
Ammorbidente	1 0 0	57.9%
Acciaio Misto	1 0 0	33.9
Libri di Testo Scolastici	1 -1 0	25%
Cartoncino Solidarietà	0 0 0	80.0%
Calzino Donna	1 -1 1	28.3%
Perizoma Donna	0 0 0	39.3%
Porcellana	1 0 0	42.2%
Spugna	1 0 0	37.6%
Shampoo	1 0 0	29.5%

Sequenze Tipiche

$< \mathbf{1,0,0} >$ (56%)

$< \mathbf{0,0,0} >$ (28%).

$\Rightarrow$ Andamento tipic

costanza.

Tipici; Individuazione e massimizzazione del TOC

	TOC_I	TOC_{EA}	TOC_{EM}	TOC_{LM}	TOC_L	$\%I$	$\%EA$	$\%EM$	$\%LM$	$\%L$
Tutti i Prodotti tranne gli stagionali										
PA(all)	0.307	0.365	0.502	0.455	0.382	5.6%	14.3%	37.0%	29.4%	13.4%
PA(3)	0.431	0.450	0.541	0.540	0.473	7.6%	14.4%	31.9%	32.4%	13.6%
AR	**0.477**	0.503	0.573	0.575	0.512	7.2%	13.2%	33.5%	32.4%	23.6%
Tutti i Prodotti										
PA(all)	0.281	0.328	0.468	0.434	0.347	7.1%	13.9%	35.3%	30.3%	13.5%
PA(3)	0.419	0.408	0.538	0.505	0.442	10.7%	10.8%	35.9%	29.2%	13.3%
AR	**0.478**	0.494	0.557	0.557	0.490	9.3%	15.1%	31.4%	31.6%	12.6%

Tabella: TOC - 5 Soglie Classiche

- Lunghezza ciclo di vita (7)
- Soglie (6)
- Categorie di prodotti (8)

Risultati più rilevanti

1. Primi acquisti ed acquisti ripetuti hanno curve identiche
2. **Gli Innovatori Tipici esistono ed hanno un TOC maggiore rispetto agli Innovatori definiti in maniera "classica"**
3. Gli Innovatori "occasionali" hanno un TOC minore rispetto agli Innovatori *heavy-user*
4. Le categorie "classiche" di Adopter non individuano il TOC massimo
5. Coerenza alta per gruppi di prodotti merceologicamente simili
6. **La disponibilità all'informazione fa aumentare il TOC**
7. Il TOC è più alto negli uomini e negli under-45