# Le conseguenze di un twitt

Linda Pagli

San Pellegrino, settembre 2013

# Twitter

- Twitter è una rete sociale che permette messaggi brevi detti tweet o cinguettii.

# Breve storia di twitter


- Rete sociale nata nel 2006.
- Gli utenti hanno gratuitamente una pagina personale, aggiornabile con messaggi brevi, max 140 caratteri detti tweet, tipo SMS.
- Architettura completamente Open Source.

# Breve storia di twitter

- Attualmente più di 500 milioni di utenti attivi, di cui 2.5 solo in Italia.
- Trecentoquaranta milioni di tweet al giorno.
- Nel 2009 diventa, a sorpresa,  anche motore di ricerca.
- 1.6 milioni di interrogazioni al giorno.
- Usato in 125 paesi.

# Come funziona

- Gli utenti hanno i loro:
  - Following: pagine delle persone seguite dall'utente.
  - Follower: persone che seguono la pagina dell'utente.
  - Twitt: i messaggi inviati

  #following, #followers and #twitt, dati associati alla pagina e ai twitt.

  Un twitt sarà tanto più importante quanti più follower di qualità saranno raggiunti

# Come funziona

- Quando si riceve un messaggio e si giudica interessante si ha la possibilità di rimbalzarlo ai propri follower. La tecnica si chiama Retweet è può essere considerato un indice di importanza del messaggio.
- Si possono marcare parole del messaggio (hashtag , come es #summerschool) per poi cercare tutti i messaggi che le contengono.

Twitter come motore di ricerca

# Perché tanto successo?

- Twitter ha cambiato il modo di fare informazione: informazione diretta da moltissime fonti, senza barriere, immediata corredata di emozioni.
- Contatto diretto (??) con persone di tutti i tipi, come grandi scrittori o artisti.

# Perché tanto successo?

- Primavera araba *
- Terremoto del Giappone e tsunami. I twitt tempestivi hanno permesso di salvare vite.
- Operazione Osama Bin laden, twittata da un pakistano che per primo si è accorto degli elicotteri USA.

* Ha fatto appena in tempo a cominciare prima che Twitter la trasformasse in un'unica enorme pubblicità di Twitter (Jonathan Franzen)

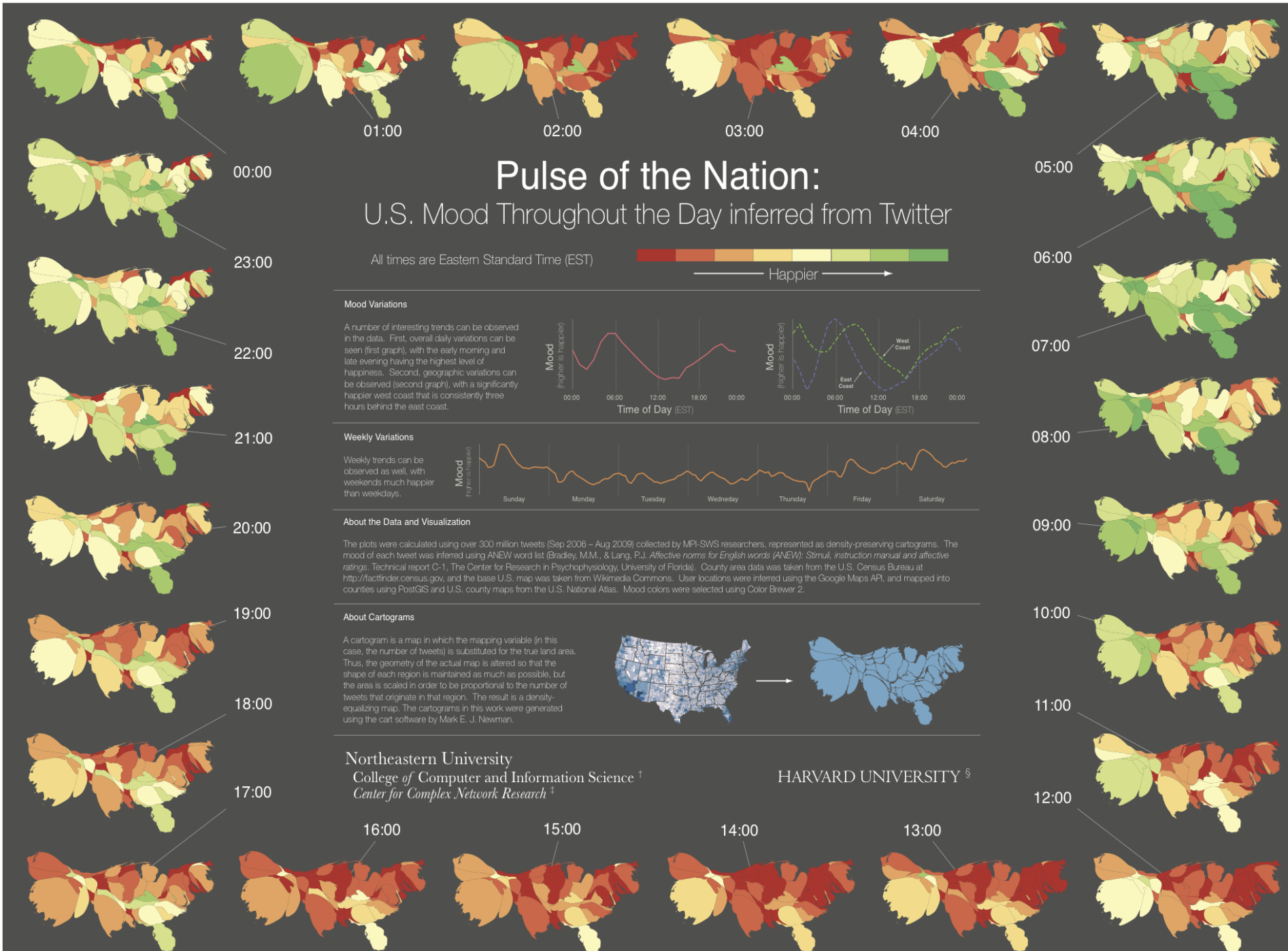
# Curiosità

- Un twitt dallo spazio; Timoty Creamer astronauta della Nasa dalla Stazione Spaziale internazionale
- Sostituisce la famosa linea rossa Obama-Putin.
- Anche il Papa twitta!

# Curiosità

http://election.twitter.com si è potuto seguire in diretta l'indice di gradimento elettorale dei due candidati alle presidenziali americane.

Sentiment Analysis sull'umore delle persone negli Stati Uniti durante le ore del giorno.




# Pulse of the Nation:

## U.S. Mood Throughout the Day inferred from Twitter

All times are Eastern Standard Time (EST)

Happier

### Mood Variations

A number of interesting trends can be observed in the data. First, overall daily variations can be seen (first graph), with the early morning and late evening having the highest level of happiness. Second, geographic variations can be observed (second graph), with a significantly happier west coast that is consistently three hours behind the east coast.




### Weekly Variations

Weekly trends can be observed as well, with weekends much happier than weekdays.




### About the Data and Visualization

The plots were calculated using over 300 million tweets (Sep 2006 – Aug 2009) collected by MPI-SWS researchers, represented as density-preserving cartograms. The mood of each tweet was inferred using ANEW word list (Bradley, M.M., & Lang, P.J. *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings*. Technical report C-1, The Center for Research in Psychophysiology, University of Florida). County area data was taken from the U.S. Census Bureau at http://factfinder.census.gov, and the base U.S. map was taken from Wikimedia Commons. User locations were inferred using the Google Maps API, and mapped into counties using PostGIS and U.S. county maps from the U.S. National Atlas. Mood colors were selected using Color Brewer 2.

### About Cartograms

A cartogram is a map in which the mapping variable (in this case, the number of tweets) is substituted for the true land area. Thus, the geometry of the actual map is altered so that the shape of each region is maintained as much as possible, but the area is scaled in order to be proportional to the number of tweets that originate in that region. The result is a density-equalizing map. The cartograms in this work were generated using the cart software by Mark E. J. Newman.



Northeastern University
College *of* Computer and Information Science †
*Center for Complex Network Research* ‡

HARVARD UNIVERSITY §

# Twitter vs Facebook

1. Guarda al presente /guarda al passato
2. Gruppi di interesse / gruppi di amici
3. Rete di informazione / rete sociale.
4. Pubblico / privato.
5. Audience mirata / universale.

# Come si finanzia

- Non ci sono pubblicità sulle pagine, e questo fa sembrare tutto più asettico, ma ci sono i twitt sponsorizzati o promozionali.
- Ci sono stati tanti finanziatori che ci hanno creduto (venture capitalist) e che vorranno guadagnare.
- Per ora la scelta di usare standard aperti ha permesso di guadagnare a tantissimi produttori di applicazioni.
- Si può anche accedere a parte dei dati e sviluppare analisi

# Ma è utile ricordare che:

Tutta l'attività fatta su internet viene registrata.
Di ogni utente si possono sapere tantissime cose….
Dalle relazioni tra gli utenti se ne possono dedurre tantissime altre!

# Ma è utile ricordare che:

- Dati : nuovo petrolio  Andrew Keen
- Data scientist: "the sexiest job of the 21-th century"  The economist
- Accesso ai dati:
  - Poco per molti
  - Moltissimo per pochi (Web giants, NSA)

# È utile ricordare che:

- Avere i dati a disposizione significa capire l'orientamento delle masse, come fanno le intercettazioni telefoniche anonime, con la ricerca di parole chiave o delle parole più usate.
- Avere un'informazione prima degli altri può fare la differenza.
- (es. un picco di follower significa un aumento immediato della popolarità)

# Esempio

- Catturare un'informazione prima di tutti puo'essere molto redditizio.
- Se una donna aspetta un figlio ha nuove necessità dall'inizio della gravidanza a molti anni dopo.
- Cambio delle abitudini
- Catturare la futura mamma da parte di un supermercato, può significare (è probabile) che comprerà lì anche il resto.

# Esempio

- Da un twitt: Sono felicissima, ho scoperto che aspetto un bambino!
- Incrocio dati: carte di credito, carte di acquisto, iscrizioni a siti, ecc. siamo tutti schedati (Gusti e preferenze comprese).
- Si inviano buoni offerta di prodotti idonei (mischiati agli altri) e se l'aggancio avviene il guadagno è sicuro!

# Altri esempi

- Sentiment anlysis in finanza: si puo' rilevare un grande interessamento a un asset .
- Si possono registrare un elevato numero di persone che si rivolgono agli ospedali e capire in anticipo lo svilupparsi di un'epidemia

# Analisi di pattern

- Quando Snowden ha rivelato che il governo americano spia l'attività su internet di tutti, Obama ha controbattuto che non si spiano singoli cittadini (cosa falsa perchè si spiano anche le mail dei singoli) ma che si cercano pattern significativi nella rete.

# Analisi di pattern

- Cercare correlazioni tra i dati

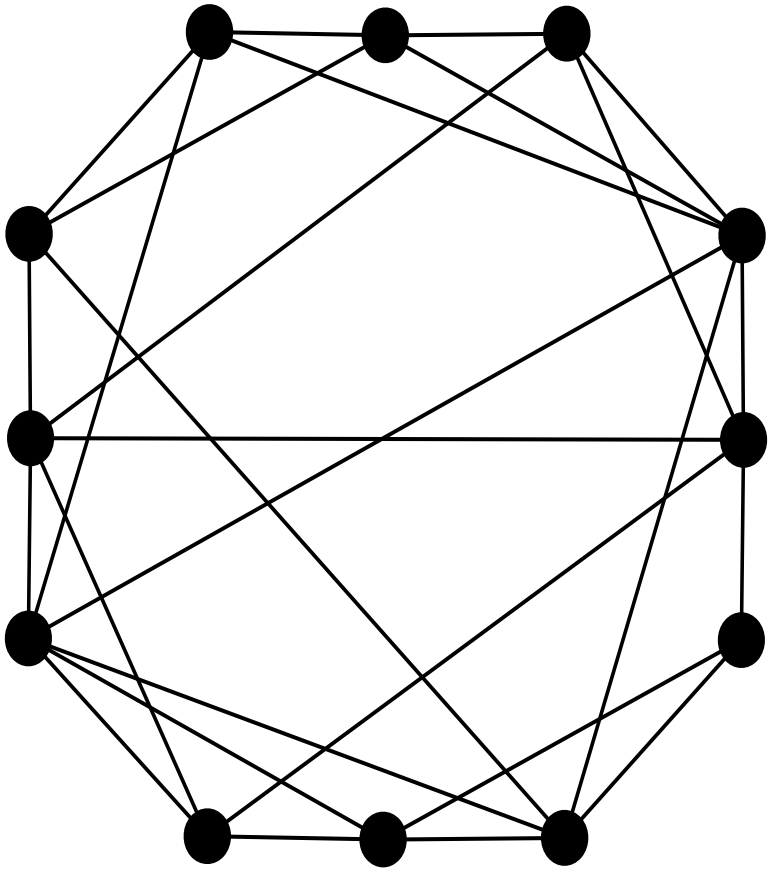
  Data Mining

- Sondare le reti del web (sociali o altre) alla ricerca di sottoreti con struttura particolare.

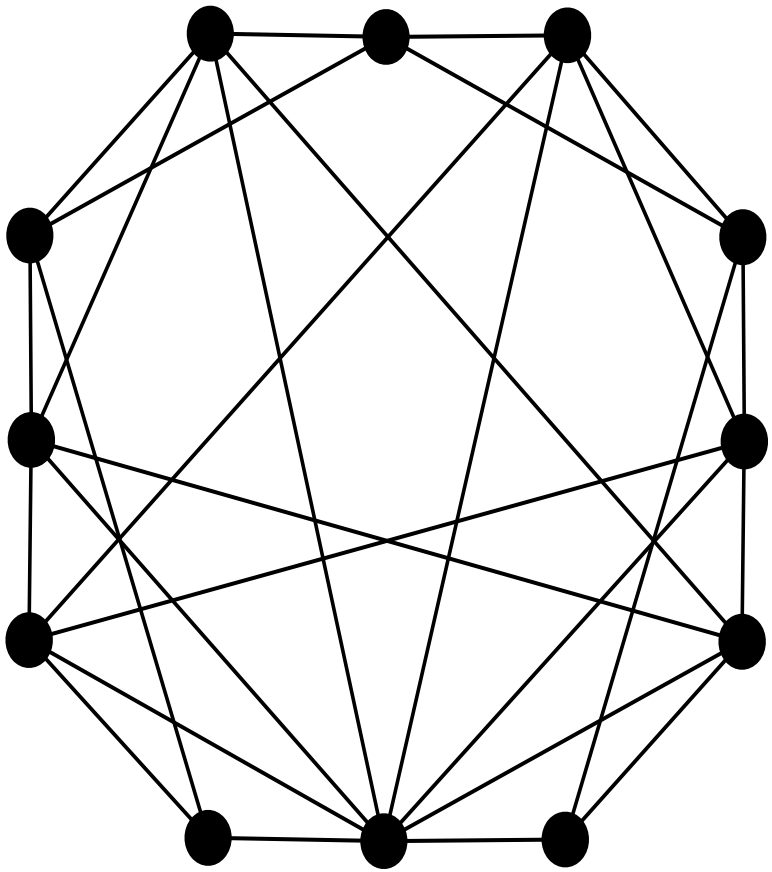
# Esempio

- Una banda di criminali decide di usare una rete sociale legale per scambiarsi materiale proibito e crittografato. Per identificarla può essere più semplice cercare se esiste una sottorete completa (clique) all'interno della rete sociale, dove ognuno scambia materiale con tutti gli altri,

# Ricerca di 4-clique in una rete di 12 nodi



no



si

# Come è possibile?

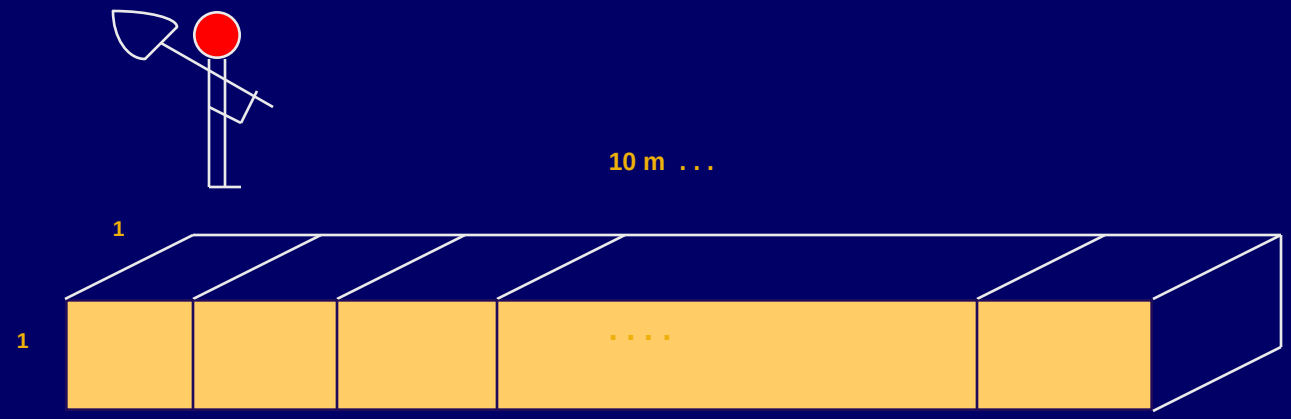
- Enorme disponibilità di potenza di calcolo unito a un uso massiccio del calcolo parallelo e distribuito.

# Il calcolo parallelo (distribuito)

Ovvero come computer eterogenei si associano per risolvere problemi troppo grandi per essere affrontati da uno solo o si accordano per stabilire strategie e modalità di comunicazione

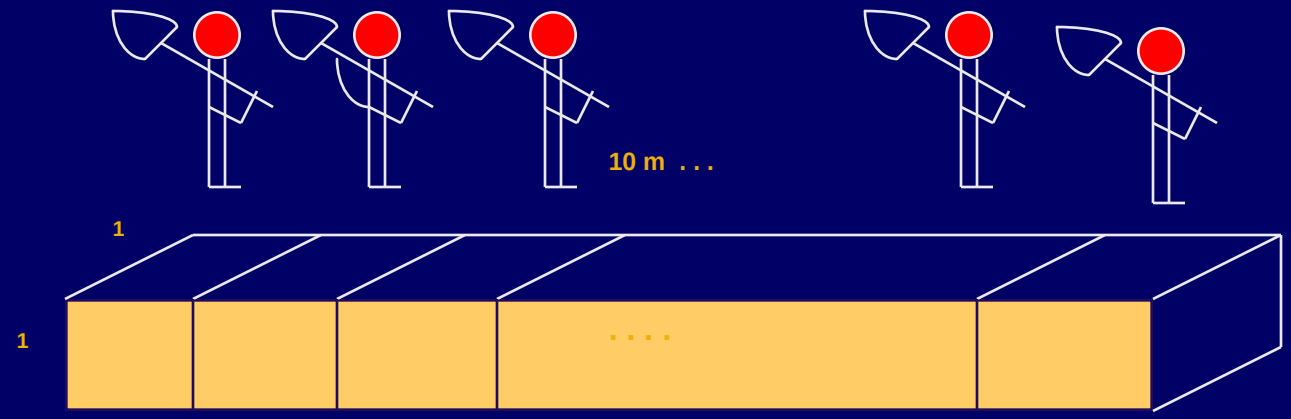
# Per scavare una buca di 10X1 m. di 1m profondità

Alla velocità oraria di 1X1X1 m:




Uno scavatore impiega 10 ore

# 10 scavatori impiegano 1 ora




Ma cosa succede se la buca deve essere profonda 10 metri?

# Entità (computer, cellulari…)

Ogni entità è in grado di

- eseguire calcoli e algoritmi con grande precisione e altissima velocità
- trasmettere messaggi

Connesse in rete

- cooperano alla soluzione di un problema
- si accordano su una strategia comune
- lavorano come un tutto unico

# Potere computazionale

La capacità intrinseca di risolvere problemi è la stessa per una singola o per un insieme di entità connesse tra loro.
Cambia radicalmente la dimensione dei problemi che è possibile affrontare e risolvere.

Esempio: Motori di ricerca, reti sociali

# Operare collettivamente è una mossa vincente?

Nascono problemi nuovi, a volte complicati a volte inaspettati….

# Problemi di deadlock
## uso di risorse in modo esclusivo

# Operare collettivamente è una mossa vincente?

Nascono problemi nuovi, a volte complicati a volte inaspettati….

Seguire l'esecuzione di un algoritmo distribuito è come cercare di seguire una conversazione animata dove tutti parlano contemporaneamente. Difficili da descrivere.

# Altre difficoltà

- Conoscenza parziale della rete
- Difficoltà a prevedere il comportamento di un algoritmo distribuito
- L'andamento del calcolo dipende dai ritardi sulle linee di comunicazione
- Un algoritmo corrisponde a numerose esecuzioni
- I parametri di valutazione sono diversi da quelli usuali

Internet: Popolarità di un protocollo

# 6 agenti devono interpretare un messaggio cifrato di 1000 pagine

1. La chiave segreta è comunicata dal comando centrale

Il messaggio può essere suddiviso in 6 parti uguali (e indipendenti) .

Impiegano 1/6 del tempo di un unico agente (+ il tempo di accordarsi sulla suddivisione del testo e la ricombinazione delle parti )

# I 6 agenti devono interpretare un messaggio cifrato di 1000 pagine

2. La chiave è lunga come una pagina.
   Per ogni pagina, eccetto la prima, la chiave è costituita dalla pagina precedente, prima della cifratura.
   Il comando centrale conosce la chiave della prima pagina, ma non conosce le successive.

Gli agenti non possono suddividersi le pagine !
Potrebbero lavorare tutti sulla stessa pagina, poi alla successiva...

- La soluzione collettiva può essere a volte necessaria,
- a volte conveniente,
- oppure inutile,
- o addirittura dannosa!

  (se le operazioni di coordinamento costano di più dell'eventuale risparmio di tempo).

# Servizi web: motori di ricerca, reti sociali



Sistemi complessi distribuiti che fanno un uso aggressivo del parallelismo.

# motori di ricerca:

- Memorizzano l'informazione nelle loro memorie private.

- Costruiscono i dizionari con le parole chiave

# motori di ricerca:

cluster di computer dislocati geograficamente in tutto il mondo.
Un cluster è composto da migliaia di Computer, semplici PC
Quelli dei motori di ricerca possono contenere repliche di tutto il web.

# Google Data Center

# motori di ricerca:

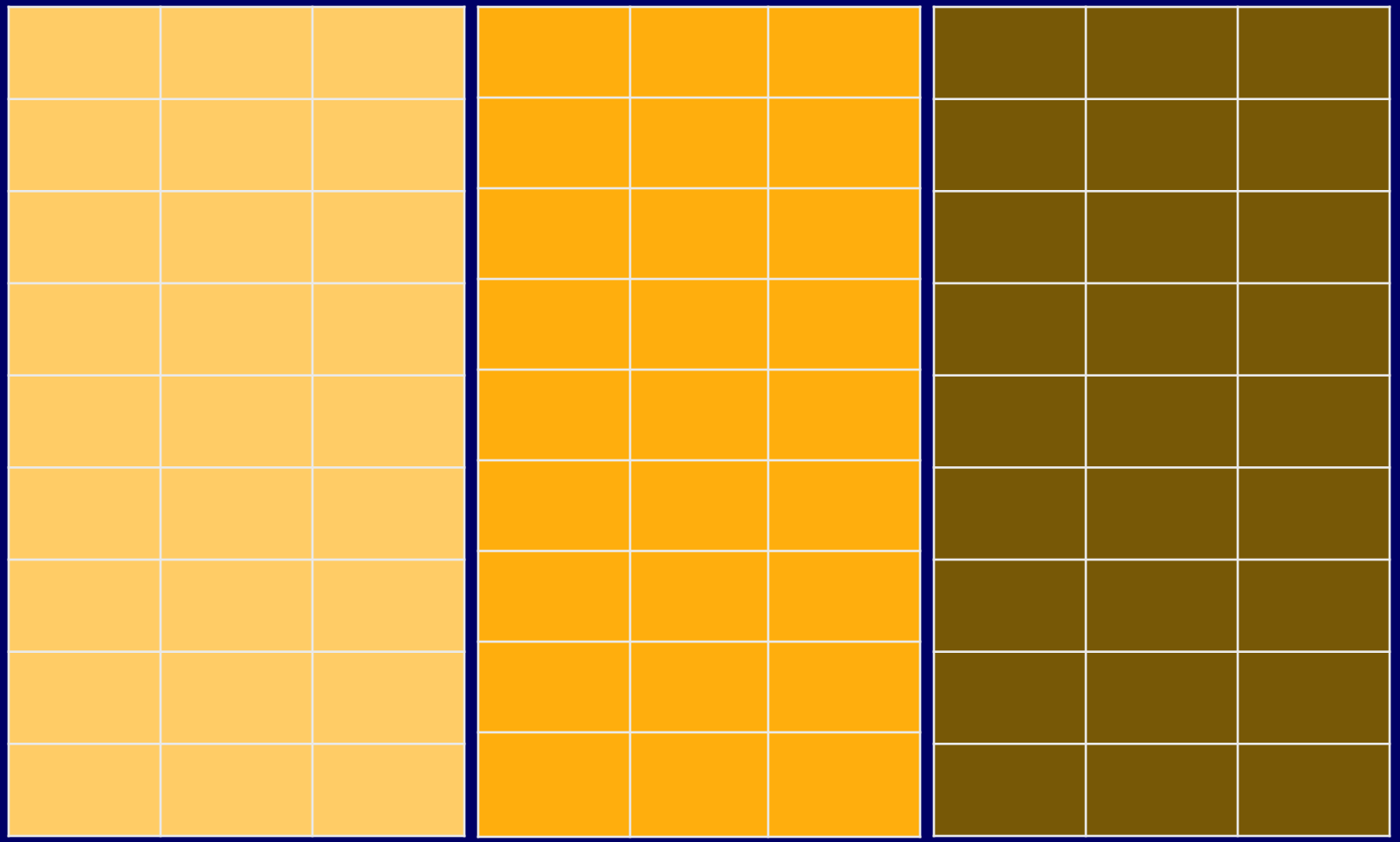
- una query in media:
  - legge centinaia di Megabytes di dati
  - consuma  decine di miliardi di cicli di CPU

- Si gestiscono milioni di queries/sec
- Una query viene smistata al cluster più vicino o, se è molto occupato, a quello più sgombro

# Come fanno a rispondere così velocemente??

L' insieme di dati da analizzare è enorme, ma i dati possono essere decomposti in pezzi di ugual dimensione.
Si esegue la distribuzione dei dati e bilanciamento del carico.
La soluzione parallela (ogni pezzo affidato a un computer diverso) è la migliore.

# Un grande array può essere suddivisa in sotto-array



- Se per ogni sotto-array si deve fare lo stesso algoritmo (es. ricerca) su dati diversi
- Se non ci sono dipendenze funzionali
- Se non è richiesta nessuna comunicazione tra i sotto-array

Abbiamo un caso ideale di computazione parallela!

# Map-Reduce

Tecnica di programmazione parallela sviluppata da Google e usata anche da Twitter per analizzare grandissimi insiemi di dati, troppo grandi per un solo computer.
Implica che il calcolo sia lo stesso per tutti i sottoinsiemi di dati diversi.
Libera il programmatore dai dettagli di della parallelizzazione, distribuzione, elezione del "capo", bilanciamento del carico ecc.

# Map-Reduce

- Esempio: Conta le occorrenze di ogni parola nel testo.

Map: conta le occorrenze di ogni parola indipendentemente in ogni sottoinsieme di dati.

Reduce: somma i valori parziali di ogni parola chiave dopo aver ordinato il sottoinsieme per raggruppare tutte le occorrenze della stessa chiave

# Map-Reduce

- La divisione del file in pezzi di uguale misura, le copie del programma (Map-Reduce) da inviare a tutte le CPU coinvolte, l'ordinamento dei dati parziali, la scrittura dei dati finali è tutto fatto automaticamente all'invocazione della procedura.

# Query di parola chiave

Avviene con lo stesso meccanismo:

- Ogni pezzo fa indipendentemente la sua ricerca e produce come risultato l'indirizzo della pagina dove è stata trovata (Map).
- Si produce la graduatoria delle pagine che la contengono in base al calcolo del page-rank(Reduce) .

# Perché Map-Reduce su enormi moli di dati è così efficace?

- Il parallelismo è spinto al massimo, la ricerche avvengono indipendentemente e il lavoro di ricombinazione è trascurabile.
- Speed-up lineare!
- Le richieste sono smistate in parallelo su cluster diversi e gestite in parallelo sui pezzi del dizionario e dell'archivio completo.

# Trovare un pattern è più difficile

- Trovare una 10-clique come sottorete di una rete di 1000 nodi.

Non esiste algoritmo migliore di quello che fa tutte le prove: determina tutti i gruppi di nodi di 10 elementi e controlla che siano clique

Lista ordinata dei gruppi

(1, 2, 3, …9, 10)
(1, 2, 3, …9, 11)

.

(991, 992, 993, …, 1000)

Quanti sono?

$2.64 \times 10^{24}$

Per ogni gruppo si devono controllare tutte le coppie di nodi che sono 45.

# Problema difficile anche in parallelo

Il numero di gruppi cresce esponenzialmente con il numero di nodi della rete n quando la dimensione dlla clique (k) tende a n/2.

# La buona notizia per i dati di twitter

- Oltre all'architettura e ai programmi anche i dati sono "aperti" anche se in piccola percentuale (1%), che pero' rappresenta una grandissima mole di dati (solo twitter lo fa).
- È possibile accedere a uno "stream" e farci le nostre ricerche per diletto, per ricerche scientifiche o semplicemente per denaro!
- Non è difficile farlo!

# Chiunque abbia un account

Può accedere a:
https://stream.twitter.com/1statuses/sample.json
e scaricare un flusso continuo di twitt (1% del totale) su cui fare analisi.
Usare operazioni compatibili con Map-Reduce per ottenere velocemente i risultati.

# Esempio

Analisi automatica di twitt per selezionare quelli che possono interessare a un utente ma che lui non vede perchè disconnessi dalla sua rete di following/followers.
(studio basato su ricerca di parole chiave e valutazione (voto) del twitt)

# E come strategia seguiamo i consigli di Grisham…

- "…da questo momento in poi tu lascerai una traccia. Non dimenticarlo mai, chiunque incontri qualsiasi cosa tu faccia, rappresenterà una tua traccia. Il segreto della sopravvivenza è proprio quello di lasciare il minor numero di tracce…"

# Bibliografia per cominciare


Barroso, Dean Hoelzle: Web Search for a Planet. The Google Cluster Architecture. IEEE Micro 2003.

Fabrizio Luccio, Linda Pagli: Storia matematica della rete. Bollati Boringhieri 2007

Fabrizio luccio, Linda Pagli: Algoritmi, divinità e gente Comune, Ed. ETS 1999 e 2012.

Charles Duhigg: La forza delle abituidini. The New York Times Magazine, Stati Uniti.

Pennacchiotti, Silvestri, Vahabi, Venturini: Making your Interest following you on Twitter.
http://zola.di.unipi.it/rossano/wp-content/papercite-data/pdf/cikm12.pdf