Big and smart data: nuovi strumenti di analisi e comprensione dei fenomeni economici. Il contributo del Dipartimento di Economia

Agar Brugiavini

Big data

Sono dataset che per dimensioni o tipologie di dati vanno oltre …

Offrono un enorme potenziale di utilizzo

- I big data sono caratterizzate da una o più delle seguenti:
  - **Volume:** dati da una grande varietà di sorgenti, incluse transazioni finanziarie, social media, sensori o *machine-to-machine*. In passato ***lo storage*** sarebbe stato un problema, ma le nuove tecnologie (quali Hadoop) e la loro integrazione nei softwares più diffusi di analisi dei dati (es.: R, Matlab) ci facilitano il compito.
  - **Velocità:** i dati fluiscono ad altissima velocità e quindi ***devono essere gestiti tempestivamente***
  - **Varietà:** I dati arrivano in qualsiasi tipo di formato - da dati strutturati e numerici in database tradizionali a non strutturati come: documenti di testo, email, video, audio, dati ticker e transazioni finanziarie e commerciali.

- Alcuni considerano anche altre due dimensioni nel caratterizzare i big data:
  - **Variabilità:** L'aumento della velocità e della varietà dei dati va unito al fatto che i flussi possono essere altamente inconsistenti e con picchi periodici.
  - **Complessità.** I dati arrivano da molteplici fonti, il che rende *difficile collegare, abbinare, ripulire e trasformare i dati trasversali*. Tuttavia, *è necessario connettere e correlare le relazioni, le gerarchie e i collegamenti* se si vuole interpretare le info.

- I big data possono essere:
  - ***Fat***: grande numero di attributi/informazioni/variabili per unità di osservazione
  - ***Tall***: grande numero di osservazioni

Smart data

Il volume del dataset non è una caratteristica sufficiente per la sua utilità: i dati devono essere puliti, filtrati e preparati per il contesto dell'analisi e la ricerca.

I big dati che sono stati ripuliti e sistemati per analisi di contenuto sono chiamati **smart data**. Nel processo di trasformazione da big data a smart data alcune delle dimensioni sopra elencate possono subire trasformazioni:

- Il volume si riduce;
- La varietà può ridursi in base a processi di screening;
- Il valore informativo e l'accuratezza aumentano.

Big data: il Contributo del Dipartimento di Economia dell'Università Ca' Foscari

La scuola di commercio di Venezia, oltre alla vocazione aziendale, linguistica, giuridica ha sempre avuto anche quella **microeconomica/macroeconomica**, **della matematica della finanza, delle assicurazioni e della probabilità**, che sono caratterizzanti dei percorsi di studio del Dipartimento.

Le aree di ricerca sui temi big-data e complessità hanno sviluppato e si propongono di sviluppare sia metodologie matematiche e statistiche per l'analisi dei dati sia analisi econometriche di banche dati di interesse per la comunità scientifica e oltre.

La sfida maggiore è quella di sostenere modelli del comportamento economico (micro e macro) alla luce di queste fonti così ricche e di poter tradurre le conoscenze che derivano da questo tipo di dati in: previsioni, analisi di scenario, strumenti di analisi, strumenti per le politiche economiche.

Nel Dipartimento di Economia si usano vari big data, questo ci ha permesso di crescere in delle particolari competenze di ricerca.

Alcuni esempi.

1. Dati networks

Flussi commerciali e Flussi finanziari

In altri casi i dati sono raccolti in forma di networks. I dati raccolti dall'**Organizzazione delle Nazioni Unite** (**COMTRADE** database) sui **flussi commerciali** (Figura 1, sinistra) e dal **BIS** (Bank of International Settlement) sui **flussi finanziari** contengono i flussi tra coppie di paesi. L'analisi di questi dati consente di valutare l'impatto su ogni singolo flusso internazionale delle politiche monetarie e fiscali messe in atto dai singoli paesi. La frequenza dei dati non è elevata, ma la loro struttura di rete rappresenta un elemento di difficoltà per l'analisi.

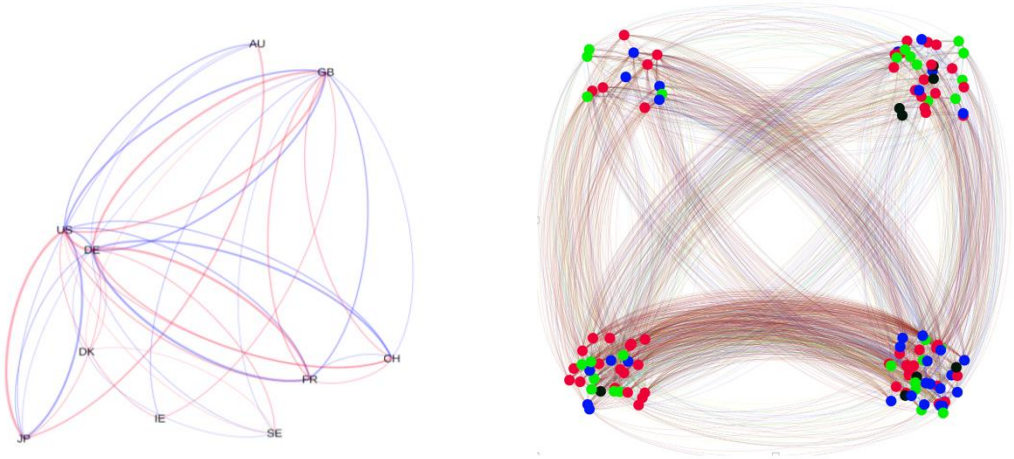

Figura 1. Sinistra: Rete Comtrade. Destra: Financial linkages.

I dati relativi alle imprese (finanziarie, assicurative e commerciali) disponibili in **Bloomberg** consentono di rilevare le relazioni tra di esse e di esprimerle in forma di networks. Il dataset delle relazioni tra imprese finanziarie quotate dal 1995 al 2016 include una sequenza di **4197 network di relazioni finanziarie con 766 nodi** (Figura 1, destra). L'analisi di questi dati consente di estrarre il livello di coesione dei mercati finanziari ed il **livello di contagio finanziario**. Gli indicatori che si ottengono sono di fondamentale importanza per le decisione di politica economica sia dei singoli stati che delle autorità sovranazionali.

2. Dati a struttura complessa

Il dataset di **17,413 piccole e medie imprese del Veneto** disponibile in **AIDA** con 14,125,362 relazioni di compartecipazione nei CDA (fenomeno di *interlocking directorate*, Figura 2), consente di studiare il comportamento delle imprese e rilevare le comunità di imprese.

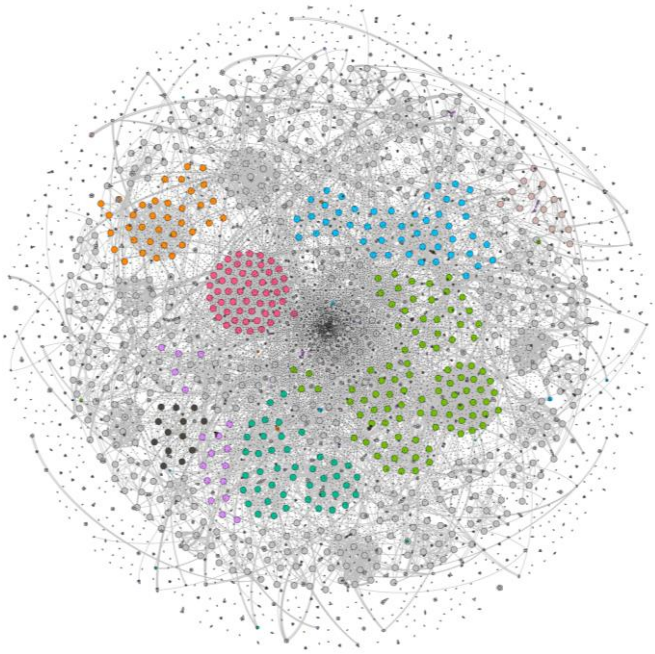

Figura 2. Interlocking directorate nel Veneto.

Le **opinioni** degli esperti di analisi economica (come le **Survey of Professional Forecasters** della **Federal Reserve Bank** of Philadelphia) consentono di analizzare le aspettative su alcune variabili rilevanti per le scelte di politica economica come l'inflazione ed il PIL ed il livello di incertezza nelle aspettative degli agenti economici. I dati sono raccolti su panel di esperti in forma di probabilità di accadimento di possibili scenari futuri. Anche in questo caso, la struttura complessa dei dati rappresenta una barriera alla estrazione delle informazioni rilevanti.

3. Dati non strutturati e "Data augmenting"

Text-Mining

Applicazioni di linguistica computazionale e **text-mining** richiedono l'esame automatico di enormi moli testi per scoprire tendenze economiche e come queste si siano incarnate in parole (es. come si è evoluta la percezione della crisi bancaria veneta e come il linguaggio sia cambiato nel corso degli anni attraverso l'analisi di migliaia di articoli pubblicati sulla stampa). Dal punto di vista tecnico ci sono grandi problemi legati sia al grande numero di articoli sia alla difficoltà di separare i testi che parlano del problema in oggetto. Due esempi di text-mining riguardano l'analisi dei **forum** e dei **social networks**.

Un esempio è rappresentato dagli **800,000 messaggi** su titoli azionari quotati, scambiati tra i **7,500 investitori** aderenti al ***forum Finanzaonline.com***, a frequenza infra-giornaliera. L'analisi dinamica della struttura di relazione tra gli utenti del forum (Figura 3) e dei contenuti dei messaggi del forum consente di determinare le reazioni collettive degli utenti del forum al livello di incertezza sui mercati azionari misurabile grazie alle banche dati finanziarie (e.g. Bloomberg). Gli investitori più esperti diventano più influenti nei momenti di elevate incertezza e la comunicazioni su piattaforme elettroniche possono avere un ruolo rilevante nel determinazione della volatilità nei mercati finanziari.

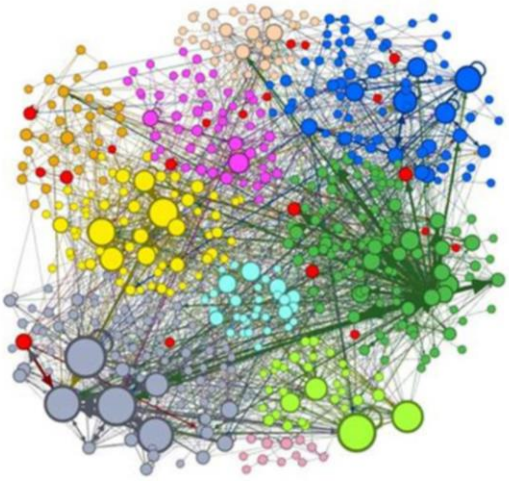

Figura 3. Scambio di messaggi tra gli utenti del forum Finanzaonline in un dato istante di tempo.

L'analisi dei messaggi **Twitter** delle rilevanti testate giornalistiche internazionali specializzate in economia e finanza (The Economist and Financial Times, e Wall Street Journal) rappresenta un altro esempio di analisi testuale. L'analisi di tutti i tweets consente per esempio di costruire nuovi indicatori a frequenza giornaliera sullo stato dell'economia e di **sentiment** degli operatori economici.

Dati amministrativi

I pazienti hanno una pluralità di contatti con diversi punti di erogazione delle prestazioni all'interno del **Servizio Sanitario Nazionale**. Per poter valutare l'impatto delle politiche sanitarie messe in atto a livello locale e nazionale è cruciale riuscire a mappare le traiettorie individuali in modo da quantificare le diverse tipologie di consumi (ricoveri ospedalieri, prestazioni specialistiche ambulatoriali, servizi di emergenza/urgenza, consumi farmaceutici), il loro grado di sostituibilità/complementarietà e il conseguente assorbimento di risorse. Questo richiede di costruire link tra differenti banche dati che raccolgono i diversi flussi informativi per ottenere dataset che registrino tutti i contatti rilevanti tra paziente e sistema sanitario.

Altri dati amministrativi: siamo in procinto di avviare una collaborazione con l'INPS per legare i dati dell'indagine SHARE con i dati sui profili delle **carriere lavorative degli intervistati**

La sfida: produzione, analisi, interpretazione e utilizzo dei big-data

1. Modelli e metodologie statistico/probabilistiche/econometriche

Sono stati sviluppati modelli matematici adeguati per l'analisi di dati a struttura complessa. Si tratta di modelli econometrici che hanno i loro fondamenti nella **teoria dei grafi** e nell'**algebra dei tensori** e che consento di catturare le caratteristiche dinamiche e di eterogeneità spaziale dei fenomeni economici analizzati.
Sono state sviluppate tecniche di inferenza statistica adeguate per l'analisi empirica dei nuovi modelli econometrici per dati con struttura di rete. Le tecniche statistiche fondano sulla teorie matematiche della **probabilità**, dell'**inferenza Bayesiana** e sulle tecniche di **Machine Learning**.

I modelli tensoriale in combinazione con le tecniche di inferenza, consentono di risolvere il problema concreto della riduzione della dimensionalità dei dati o dei modelli in modo simile a quanto accade nell'analisi fattoriale. Altri problemi pratici che questi modelli consento di risolvere sono il filtraggio delle informazioni più rilevanti per i modelli economici e l'elaborazione di early warning systems per il supporto alle decisioni dei policy makers.

2. Gli strumenti di elaborazione:

Inaugurato nel 2014 su iniziativa del Dipartimento dei Economia e del DAIS, il **Sistema per Calcolo Scientifico di Ca' Foscari** (SCSCF) è un insieme di elaboratori e programmi che consentono di soddisfare le crescenti esigenze dei ricercatori del Dipartimento di Economia in tema di calcolo parallelo ed intensivo in termini di potenza di calcolo di archiviazione di grandi moli di dati. Il sistema include 7 nodi di calcolo ed un nodo per applicazioni Big Data, per un totale di 260 cores.

3. La costruzione dei modelli economici e la produzione di Big-Data

Una sfida ancora più accentuata è la possibilità di analizzare i big-data per rispondere a temi concreti di economia e politica economica (in generale le scienze sociali).

Il Dipartimento eccelle non solo nella parte di analisi dei dati già descritta ma anche nella produzione dei dati stessi.

- **SHARE**
  La Rilevazione SHARE si configura come "Long Data" ma anche Big Data (oltre 300mila individui intervistati ogni 2 anni dal 2004 in 28 paesi europei – incluso Israele).
  Circa 300 variabili (caratteristiche) rilevate + metadati (ticks delle riposte – pause degli intervistatori).
  Dimensione SHARELIFE ricostruisce i dati all'indietro per conoscere gli effetti delle politiche nei primi anni di vita.

Più recentemente gli intervistati dovranno indossare un “accellerometro” per registrare il loro livello di attività nell’arco di una settimana.

In questo caso i temi di ricerca in ambito economico-sociale **motivano e guidano la creazione di BIG-DATA.**
Ad esempio: anni addizionali di scuola dell’obbligo hanno effetti positivi (negativi) nel lungo periodo sui tassi di occupazione e salari? Teoria del capitale Umano. Ci sono “scarring effects” dovuti a esperienze negative macroeconomiche (carestie, guerre, recessioni) o microeconomiche (disoccupazione) o familiari (maltrattamenti da bambini)