BINet
Business Intelligence & Network Analysis

La **sfida** della Social Network Analysis nel Disegno di Algoritmi

Marco Pellegrini
Istituto di Informatica e Telematica del CNR

Contesto (I)

- Il termine "**rete**" ("rete sociale", rete complessa") e' pesantemente **sovraccarico di significati**, in quanto ricorrente in molteplici campi applicativi (bibliometria, biologia, sociologia, comunicazioni, ..)

- I problemi (e gli algoritmi che li risolvono) relativi alle reti tuttavia vengono spesso formalizzati tramita le "**teoria dei grafi**" che unifica gli **aspetti stutturali** che emergono nei vari campi applicativi.

Contesto (II)

- Una **tradizione di analisi di reti sociali**, che data almeno dagli anni ’30, e’ stata di recente potenziata grazie a:
  - L’affermarsi di strumenti informatici per la raccolta di **grosse moli di dati** a basso costo.
  - **Convergenza** tra reti informatiche (Internet, WWW) e reti sociali.

Contesto (III)

- Scoperta di **caratteristiche comuni** in molti campi applicativi diversi:
  - Distribuzioni “power-law” (distribuzione di Pareto)
  - Fenomeni “small-world”
- Applicazione di metodi di **meccanica statistica** (dalla fisica) e di **teoria degli algoritmi** (dall’informatica).

Diversi livelli di analisi (I)

Analisi a livello dei singoli elementi

- “**Quale e’ l’elemento piu’ importante?** “,
- **“Quanto importante e’ un dato elemento?”**

- Collettivamente queste valutazioni numeriche (o in rango) vengono chiamati “**Indici di centralita’**”

- Vi sono svariati indici che si applicano a **diversi tipi** di reti e/o a **diverse nozioni** di “importanza”.

Diversi livelli di analisi (I)

Analisi a livello dei singoli elementi

- L'esempio piu' semplice e' il "**Degree centrality**" dove la centralita' di un nodo e' data semplicemente dal **numero di archi incidenti** sul nodo (entranti e/o uscenti) .

- L'esempio piu' noto e' **Pagerank di Google**, che assegna un valore d'importanza ad una pagina web proporzionale alla **probabilita' di terminare una visita "random walk"** del grafo del web su quella pagina.

Diversi livelli di analisi (II)

Analisi a livello di gruppi di elementi

- I gruppi possono esser definiti in base al **numero** (**o forza**) **delle connessioni reciproche**, o in base alla presenza di **schemi specifici** (**patterns**).

- Un classico problema di "**clustering**" di un grafo: trovare una "**partizione**" dei nodi in modo che per ogni modulo **i legami interni siano piu' forti di quelli esterni**.

- Ma e' difficile indovinare il **numero di moduli** giusto.

Diversi livelli di analisi (II)

Analisi a livello di gruppi di elementi

- **Componenti dense**: non e' necessario associare ogni nodo ad un modulo. Un nodo puo' appartenere a piu' moduli (sovrapposizione)

- **Componenti dense** : --> significato specifico cambia per la specifica rete (**interpretazione**)

- **Componenti dense** : --> comuni **proprieta' strutturali**: piccolo diametro, robustezza rispetto a cancellazione di archi, fortemente connesso (**ricerca**).

Diversi livelli di analisi (III)

Analisi a livello di rete globale

- Identificazione di **distribuzioni caratteristiche** (power-laws)

- Identificazione di **indici caratterisitci** che discriminano reti sociali dagli altri tipi di reti.

- **Modelli generativi** per reti sociali

- Distribuzione di **sottografi caratteristici**

Specificita' delle reti sociali (I)

- Reti sociali esibiscono una **correlazione positiva** tra tra il grado di nodi vicini. Ossia, nodi con grado alto sono vicini a nodi con grado alto, nodi con grado basso sono vicini a nodi con grado basso.

- Altri tipi di reti (tecnologiche, biologiche, etc..) hanno correlazione nulla o negativa.

- Il coeff. di correlazione e' **0** per *reti random*, **1** per *reti perfettamente assortative*, **-1** in *reti perfetametne anti-assortative*.

Specificita' delle reti sociali (I)

Tipo	Classe	n	m	correlaz.
Sociale	Attori di films	450K	25.5M	0.200
Sociale	Direttori di societa'	7K	56K	0.276
Sociale	Coautori di articoli	52K	245K	0.363
Comunic.	Internet	10K	32K	-0.189
Biologico	Rete metabolica	0.7K	3.6K	-0.240

Specificita' delle reti sociali (II)

- Le reti sociali hanno **un'altra transitivita'** (che le differerenzia da reti random con la stesso profilo.

- La transitivita' e' la propensione di due nodi con molti vicini in comune ad avere un legame tra loro.

- Per molte **reti non-sociali** la transitivita' e' vicina a quella di **reti random** (con lo stesso profilo).

- Misurato dal **coefficiente di clustering** C = 3 x (numero triangoli)/ (numero triplette connesse). Un numero tra 0 ed 1.

Specificita' delle reti sociali (II)

Tipo	Classe	n	m	C
Sociale	Attori di films	450K	25.5M	0.20
Sociale	Direttori di societa'	7K	56K	0.59
Sociale	Coautori di articoli	52K	245K	0.45
Comunic.	Internet	10K	32K	0.035
Biologico	Rete metabolica	0.7K	3.6K	0.09

La **sfida** della complessita' per la SNA

- Un grafo G=(V,E) e' caratterizzato da un insieme V di **n** nodi ed un insieme E di **m** archi.

- Tipicamente un algoritmo su G sviluppato secondo la teoria dei grafi e' efficiente in termini di caso pessimo solo **sui parametri m ed n.**

- Tuttavia **raramente** gli algoritmi standard sfruttano le caratterisiche delle SN (assortativita' e transitivita') per migliorare i tempi di calcolo.

- Molti metodi usano i **modelli generativi** delle SN.

La **sfida** della complessita' per la SNA

• Esempio: per il calcolo dell'indice di centralita' "**Betweenness**" basato sul numero di cammini minimi che attraversano un dato nodo, l'algoritmo migliore (**Brandes, 2001**) ha tempo **O(nm)**.

•Su di un grafo di Sociali di Sistemi Territoriali SrL di 0.5M nodi, 1.1M di archi una workstation impiega circa **4 mesi** di tempo di calcolo.

• L'approccio da noi sviluppato che sfrutta la presenza di molti nodi vicini a basso grado impiega circa **2,5 giorni**.

La **sfida** della complessita' per la SNA

- Per la determinazione dei **gruppi fortemente connessi** di nodi un metodo classico di **Girvan e Newman 2002** usa tempo **$O(nm^2)$**.

1) Calcola il valore di "betweenness" degli archi
2) Rimuovi l'arco con "betweenness" piu' alta
3) Analizza la "modularita'" delle componenti connesse del grafo residuo
4) Ritorna al passo 1) se vi sono ancora archi.
5) Seleziona la decomposizione generata con modularita' maggiore.

La sfida della complessita' per la SNA

- Problema (a): troppo lento per grafi di grosse dimensioni
- Problema (b): ogni nodo deve essere associato ad un gruppo connesso
- Problem (c): usa una nozione relativa e non locale di densita'.

L'algoritmo GN **non sfrutta** esplicitamente l'alta "transitivita'" delle reti sociali, ossia l'alto numero di triangoli nelle componenti dense.

La sfida della complessita' per la SNA

Sul grafo di Sistemi Territoriali
Nodi: **0.5M**
Archi **1.1M**

Toviamo:
Numero gruppi densi: 1248
Dimensione: da 12 a 5 nodi
Densita' dal 50% a' 100%

In circa **6 secondi.**
Copertura del grafo circa l' 1%.

Core number dei nodi

- Il valore massimo della **core decompostion** e' spesso un numero drammaticamente piu' piccolo del grado massimo.

Sia nei gafi sociali reali che nei modelli standard.

Modello	n	m	D ave	D max	C max
Erdos-Renyi	100K	1M	20	41	14
Barabasi-Albert	50K	100K	4	642	3
Weibull	100K	300K	6.15	377	9
Pareto	97K	293K	6.04	938	22

Conclusioni

- Reti sociali **sempre piu' grandi** (in termini di numero di entita' e relazioni) e **complesse** (in termini di fenomeni che rappresentano) vengono prodotte come risutlato di "tracce" nei sistemi informatici.

- **Algoritmi generici** che non sfruttino a fondo le caratteristiche struturali delle reti sociali sono **troppo lenti** per poter operare su dati di grandi dimensioni.

Conclusioni

- Nel progetto BINet si sono sviluppati **algoritmi specifici per Reti Sociali** per due problemi: Calcolo dell'indice di centralita' "betweennness", ed in calcolo di sottografi densi.

- L'efficienza computazionale puo' essere sfruttata per:
  (1) **migliorare l'interattivita'** tra l'utente (analista) e lo strumento per la visualizzazione dei dati,
  (2) permettere **analisi di "secondo livello"** in tempo reale.