Web-communities e cut-communities

1.1 Comunità

Lo studio sperimentale delle reti (sociali) di grandi dimensioni evidenzia l'esistenza di regioni altamente interconnesse, ossia, di regioni che hanno un'alta concentrazione di archi al loro interno e che, di contro, la densità degli archi che connettono tali regioni le une con le altre è sensibilmente più bassa. Questa caratteristica delle reti reali è chiamata *struttura a comunità* o *clustering*.

Sono state proposte, in letteratura, numerose definizioni formali di queste regioni coese di individui, ciascuna in grado di catturare alcune delle caratteristiche di coesione sulle quali si voleva indagare.

Una definizione di "regione altamente coesa" particolarmente naturale è quella che considera la relazione fra il numero di legami che ciascun individuo ha all'interno della regione cui appartiene e il numero di legami che egli ha esternamente ad essa. In altri termini, informalmente, una regione è una *web-community* se ciascun individuo appartenente ad essa è maggiormente connesso agli altri individui nella stessa regione che ad individui esterni alla regione.
Formalmente, dato un grafo $G = V,E)$ e, per $u \in V$, detto $N(u)$ l'insieme degli individui coi quali l'individuo u ha una relazione, ossia, $N(u) = \{v \in V : (u,v) \in E$, una web-community è definita come segue.

Definizione 1.1 (Web-communities): *Dato un grafo $G = (V,E)$, una* web-community *è un sottoinsieme (proprio) dell'insieme dei nodi $C \subset V$ tale che*

$$\forall\, u \in C\ [\ |N(u) \cap C|\ >\ |N(u) - C|\ ].$$

La precedente definizione può essere rilassata nella seguente definizione di *weak web-community*.

Definizione 1.2 (Weak web-communities): *Dato un grafo $G = (V,E)$, una* weak web-community *è un sottoinsieme (proprio) dell'insieme dei nodi $C \subset V$ tale che*

$$\forall\, u \in C\ [\ |N(u) \cap C|\ \geq\ |N(u) - C|\ ].$$

D'altro canto, osserviamo, se una comunità presenta un'alta concentrazione di archi quando rapportata alla concentrazione di archi che la connettono all'esterno, è evidente che un insieme di nodi è una comunità se per isolare i nodi che la costituiscono dal resto del grafo è necessario eliminare un numero ridotto di archi. Appare, dunque, anche naturale la seguente definizione di *cut-community*.

Definizione 1.3 (Cut-communities): *Dato un grafo $G = (V,E)$, una* cut-community *rispetto alla coppia $\langle s,t \rangle$ è un sottoinsieme (proprio) dell'insieme dei nodi $C \subset V$ tale che $s \in C$, $t \in V - C$ e*

$$|\{(u,v) \in E\ : u \in C\ \wedge v \in V - C\ \}|\ =\ \min\ \{|\{(u,v) \in E\ : u \in C'\ \wedge v \in V - C'\ \}|\ :\ C' \subset V\ \wedge\ s \in C'\ \wedge\ t \notin C'\ \}.$$

1.2 Partizonare in comunità

La suddivisione in comunità di una rete può avere diverse applicazioni concrete; ad esempio partizionare in cluster i *web client* che hanno interessi simili e sono geograficamente vicini tra loro può migliorare la qualità di un servizio fornito sul *World Wide Web*, dove ogni insieme di clienti potrebbe essere gestito da un server dedicato. Inoltre, identificare comunità di consumatori aventi interessi simili in una rete permette di mettere in relazione utenti e prodotti in modo da creare un efficacie sistema di raccomandazioni. In generale, identificare i diversi moduli e i loro confini all'interno di una rete permette una classificazione dei nodi in accordo con la loro posizione strutturale all'interno dei moduli. Inoltre, la ricerca di comunità all'interno di un grafo è utile anche per identificare i moduli che lo compongono e, possibilmente, la sua organizzazione gerarchica, usando solo le informazioni codificate nella topologia del grafo.

Occupiamoci, dunque, della questione di individuare le comunità in un grafo. Più precisamente, studiamo in questo paragrafo il problema decisionale che consiste nel decidere se un dato grafo è partizionabile in due comunità (in accordo ad una qualche definizione di comunità).

Ricordiamo che un *taglio* in un grafo è una partizione dei nodi del grafo in due sottoinsiemi non vuoti e che la *misura del taglio* è in numero di archi che hanno un estremo in ciascuno dei due sottoinsiemi della partizione.
Osserviamo subito che una cut-community è individuata (per definizione) da un taglio di misura minima nel grafo e che, viceversa, ogni taglio di misura minima all'interno del grafo partiziona il grafo in due cut-communities. Poiché esistono algoritmi in grado di individuare il taglio minimo in un grafo in tempo polinomiale, possiamo collocare il problema di decidere se un grafo è partizionabile in due cut-communities nella classe **P**. Tuttavia, la partizione individuata da un taglio minimo non sempre individua comunità significative, come illustrato nell'esempio seguente.

Esempio 1.1: Sia $G = (V,E)$ il grafo in cui l'insieme V è diviso in due insiemi A e B tali che $A \cup B = V$ e $A \cap B = \emptyset$ così definiti:

- A è un grafo completo di cinque nodi, u_0, u_1, u_2, u_3, u_4.
- B è un grafo completo di cinque nodi, v_0, v_1, v_2, v_3, v_4.

Il grafo G è ottenuto unendo A e B tramite gli archi (u_i, v_i), (u_i, v_{i+1}), (u_i, v_{i+2}) (le somme sono modulo 5).
È facile verificare che i tagli minimi in G sono tutti e soli i tagli che isolano singoli nodi, che hanno cardinalità 7. D'altra parte, pare poco ragionevole considerare comunità un singolo nodo.

Comunque, ove il taglio minimo produca una partizione in comunità significative, le due cut-communities individuate dal taglio sono anche weak web-communities, come mostrato nel teorema seguente.

Teorema 1.1: *Siano $G = (V,E)$ un grafo e $C \subset V$ una cut-community per G tale che $|C| > 1$ e $|V - C| > 1$. Allora,*

$\forall u \in C \; [\; |N(u) \cap C| \;\geq\; |N(u) - C| \;]$, *e*

$\forall u \in V - C \; [\; |N(u) - C| \;\geq\; |N(u) \cap C| \;]$.

Dimostrazione: Sia C una cut-community per G tale che $|C| > 1$. Supponiamo per assurdo che C non sia una weak web-community; in questo caso deve esistere un nodo $u \in C$ tale che

$$|N(u) \cap C| \;<\; |N(u) - C|.$$

Poiché $|C| > 1$, allora C contiene almeno un altro nodo v, distinto da u, ossia, $C - \{u\} \neq \emptyset$. Dunque, i due sottoinsiemi $C - \{u\}$ e $(V - C) \cup \{u\}$ sono un taglio del grafo; inoltre la misura di questo nuovo taglio è

$$\begin{aligned} |\{(x,y) \in E : x \in C - \{u\} \;\wedge y \in (V - C) \cup \{u\} \;\}| &= |\{(x,y) \in E : x \in C \;\wedge y \in V - C\}| + |N(u) \cap C| - |N(u) - C| \\ &< |\{(x,y) \in E : x \in C \;\wedge y \in V - C\}|. \end{aligned}$$

Ossia, la misura del taglio $\langle C - \{u\}, (V - C) \cup \{u\} \rangle$ è minore della misura del taglio $\langle C, V - C \rangle$, contraddicendo l'ipotesi che C fosse una cut-community e che, dunque, il taglio $\langle C, V - C \rangle$ fosse un taglio di misura minima.

Analogamente si dimostra che $V - C$ è una weak web-community. □

D'altra parte non sono noti algoritmi polinomiali in grado di trovare tagli minimi in cui ciascun sottoinsieme del taglio contenga almeno due nodi. In effetti, appare poco probabile che sia possibile partizionare un grafo in due web-communities in tempo polinomiale, come dimostrato formalmente nel Teorema 1.2.
Prima di procedere con la dimostrazione formale, abbiamo bisogno di premettere un risultato preliminare.

Lemma 1.1: *Sia $G = (V,E)$ un grafo contenente un nodo u di grado 2; se G è partizionabile in due web-communities allora per ogni partizione di G in due web-communities $\langle C, V - C\rangle$, u e i due nodi adiacenti a u sono nella stessa comunità.*

Dimostrazione: Siano (u,x) e (u,y) i due archi incidenti su u; se x e y fossero contenuti in due comunità diverse, diciamo $x \in C$ e $y \in V - C$, allora si avrebbe

$$|N(u) \cup C| = |N(u) - C|$$

e, dunque, u non potrebbe appartenere né a C né a $V - C$.

Dunque, x e y devono appartenere alla stessa comunità e, conseguentemente, anche u deve appartenere ad essa. □

Teorema 1.2: *Decidere se un dato grafo $G = (V,E)$ può essere partizionato in due web-communities è un problema **NP**-completo.*

Dimostrazione: Il problema è in **NP**: infatti, un certificato per una istanza $G = (V,E)$ è un insieme $C \subset V$ e verificare che C e V_C sono web-communities richiede, banalmente, tempo polinomiale.

Per dimostrare la completezza del problema per la classe **NP**mostriamo una riduzione polinomiale dal problema 3SAT. Sia, dunque $f = c_1 \wedge c_2 \wedge \ldots \wedge c_m$ una istanza di 3SAT, ove ciacuna clausola c_j è la disgiunzione di tre letterali appartenente all'insieme $X = \{x_1, x_2, \ldots, x_n\}$ di variabili booleane. Il grafo $G = (V,E)$ corrispondente ad f è illustrato in figura 1.2. Prima di procedere con la sua descrizione, osserviamo che

1) V contiene i due nodi T e F, che rappresentano i valori `true` e `false` e che dovranno essere contenuti in due web-communities differenti affinché G sia partizionabile in due web-communities, come vedremo a breve.

2) A ciascuna variabile x_i, $i = 1, \ldots, n$, corrisponde un sottografo individuato dai nodi x_i, $\overline{x_i}$, y_i, z_i, t_i e f_i collegati come illustrato in figura. Se x_i e $\overline{x_i}$ fossero contenuti entrambi nella stessa comunità allora anche y_i e z_i dovrebbero essere contenuti nella stessa comunità. Osserviamo, ora, che t_i e f_i hanno entrambi grado 2: questo implica, per il Lemma 1.1, che esiste una partizione in due web-communities soltanto se y_i e T sono nella stessa comunità e z_i e F sono nella stessa comunità. Di conseguenza, se x_i e $\overline{x_i}$ fossero contenuti entrambi nella stessa comunità, allora anche tutto il sottografo corrispondente alla variabile booleana x_i e T e F sarebbero nella stessa comunità.

3) A ciascuna clausola c_j, $j = 1, \ldots, m$, corrisponde un nodo c_j e i tre nodi l_{j_1}, l_{j_2} e l_{j_3}; il nodo c_j è collegato, oltre che a l_{j_1}, l_{j_2} e l_{j_3}, ad i nodi che corrispondono ai letterali che costituiscono la clausola: pertanto esso ha grado 6. Poiché l_{j_1}, l_{j_2} e l_{j_3} hanno grado 2, in virtù del Lemma 1.1 esiste una partizione di G in due web-communities solo se c_j è nella stessa comunità di T; inoltre, poiché c_j ha grado 6 e tre dei suoi vicini sono nella stessa comunità di T, affinché c_j possa essere contenuto nella stessa comunità di T senza violare i vincoli inposti dalla definizione di web community è necessario che almeno uno dei nodi corrispondenti ai suoi letterali sia nella stessa comunità di T.

4) G è completato aggiungendo, per ogni nodo x_i e $\overline{x_i}$, un opportuno numero di nuovi vicini (nuovi nodi lasciati senza nome): se il letterale x_i ($\overline{x_i}$) compare in k clausole, allora il nodo x_i ($\overline{x_i}$) viene collegato a $k+1$ nuovi nodi di grado 1. Tali nuovi nodi servono unicamente a garantire che i nodi x_i e $\overline{x_i}$ possano appartenere a qualunque delle due comunità, senza che i loro vicini altri che y_i e z_i li vincolino in alcun modo.

Dalla discussione appena condotta, risulta, quindi, che, per ogni $j = 1, \ldots, m$, il nodo c_j, deve essere nella stessa comunità di T e che, affinché questo sia possibile, almeno uno dei nodi corrispondente ad un letterale nella clausola c_j deve essere nella stessa comunità di T.

Mostriamo, ora, che se T e F sono contenuti nella stessa comunità allora quella comunità deve contenere tutti gli elementi di V: infatti, se T e F sono contenuti nella stessa comunità, allora quella stessa comunità contiene anche i

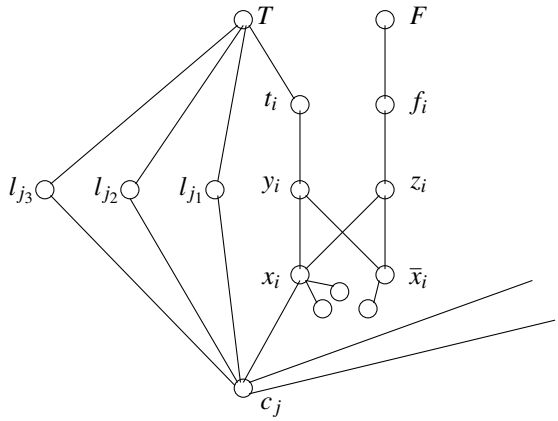

Figura 1.1: Riduzione da 3SAT al problema del partizionamento di un grafo in due web-communities.

nodi y_i e z_i, per ogni $i = 1,\ldots,n$. Dunque, se il letterale x_i (o, analogamente, $\overline{x}_i$) è contenuto in k clausole, il nodo x_i (o $\overline{x}_i$) ha $k+2$ vicini nella stessa comunità di T: i due nodi y_i e z_i e tutti i nodi corrispondenti alle clausole in cui è contenuto. Conseguentemente, i $k+1$ nodi senza nome di grado 1 ai quali è collegato non possono evitare che il nodo x_i (o $\overline{x}_i$) sia contenuto nella stessa comunità in cui sono contenuti T e F. Questo costringe tutti i nodi senza nome ad essere contenuti nella stessa comunità.
E, quindi, se T e F sono contenuti nella stessa comunità allora quella comunità deve contenere tutti gli elementi di V. O, equivalentemente, affinché G sia partizionabile in due web-communities è necessario che T e F siano contenuti in due comunità differenti.

Riassumendo: affinché G sia partizionabile in due web-communities è necessario che T e F siano contenuti in due comunità differenti, e se T e F sono in due comunità differenti, per quanto discusso al punto 2), per ogni $i = 1,\ldots,n$, anche x_i e $\overline{x}_i$ sono in due comunità differenti (uno di loro nella stessa comunità di T, l'altro nella stessa comunità di F). Questo significa che em ogni partizione di G in due web-communities corrisponde ad una assegnazione di verità per le variabili in X.

D'altra parte, per quanto discusso al punto 3), la partizione di G in due comunità è possibile solo se, per ogni $j = 1,\ldots,m$, almeno uno dei nodi corrispondenti ad un letterale contenuto nella clausola c_j è nella stessa comunità di T.

In conclusione, se G è partizionabile in due comunità allora possiamo assegnare il valore `vero` a tutti i letterali corrispondenti a nodi che sono nella stessa comunità di T (e `falso` a tutti gli altri). Poiché ogni clausola ha un letterale vero (per quanto appena osservato), ogni clausola è soddisfatta da questa assegnazione di verità. E, quindi, f è soddisfatta da questa assegnazione.
Viceversa, se esiste una assgnazione di verità a che soddisfa f allora inseriamo nella stessa comunità di T tutti i nodi x_i tali che $a(x_i) =$ `vero` e tutti i nodi $\overline{x}_i$ tali che $a(x_i) =$ `falso`. Poiché a soddisfa tutte le clausole in f, allora, per ogni $j = 1,\ldots,m$, il nodo c_j ha 4 dei suoi 7 vicini nella stessa comunità di T (i tre nodi l_{j_1}, l_{j_2} e l_{j_3}, oltre al nodo corrispondente al suo letterale `vero`): dunque, per ogni $j = 1,\ldots,m$, il nodo c_j può essere effettivamente inserito nella stessa comunità di T, come richiesto dal Lemma 1.1. Questo prova che a induce una partizione di G in due comunità.

La dimostrazione è conclusa osservando che G viene costruito in tempo polinomiale nella dimensione di f e X. □

In effetti, rilassare la definizione di web-community in quella di weak web-community non modifica la complessità computazionale del problema del partizionamento di un grafo, come affermato nel seguente teorema del quale omettiamo la dimostrazione.

Teorema 1.3: *Decidere se un dato grafo $G = (V,E)$ può essere partizionato in due weak web-communities è un problema **NP**-completo.*