Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Reti Multilayer: proprietà e metodi per l'identificazione di comunità

Annalisa Socievole, Sabrina Celia

RT-ICAR-CS-19-09 **Dicembre 2019**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci 8-9C, 87036 Rende, Italy, URL: *www.icar.cnr.it*
– Sezione di Napoli, Via P. Castellino 111, 80131 Napoli, URL: *www.icar.cnr.it*
– Sezione di Palermo, Via Ugo La Malfa, 153, 90146 Palermo, URL: *www.icar.cnr.it*

Indice

1 ABSTRACT

Molti sistemi del mondo reale sono costituiti da sottosistemi multipli con differenti livelli di interconnettivitá. É importante tener conto di questi aspetti per comprendere meglio l'organizzazione di tali sistemi. É quindi necessario generalizzare la teoria delle reti tradizionali, che tiene conto solo di un tipo di relazione tra gli oggetti costituenti un sistema, a modelli e strumenti che permettano lo studio di sistemi multilivello in maniera adeguata. Nel presente rapporto vengono introdotti i concetti base della teoria delle reti multilayer ed alcune misure base giá note nel caso singolo layer opportunamente estese. Inoltre saranno descritti gli algoritmi proposti negli ultimi anni per scoprire la struttura a comunitá in questo tipo di reti multilayer.

2 INTRODUZIONE

2.1 SCOPO DEL DOCUMENTO

Sistemi del mondo reale possono essere rappresentati da reti complesse che descrivono gli oggetti costituenti il sistema e le interconnessioni tra essi. Le reti possono essere analizzate a diversi livelli di granularitá. Il livello dei nodi é la scala piú piccola da studiare. A questo livello il grado di un nodo puó dare informazioni preziose sul ruolo svolto dagli oggetti che partecipano alla rete. Piú interessante é il livello delle comunitá in cui si studia la suddivisione di una rete in gruppi aventi intra-connessioni dense, e inter-connessioni sparse. Tale livello fornisce una descrizione mesoscopica di una rete in cui gli elementi sono le comunitá e non i nodi. Questa suddivisione é tipica di molte reti.

Una comunitá del mondo reale é un gruppo di individui con interessi e/o caratteristiche economiche, sociali o politiche comuni, che spesso vivono relativamente nelle immediate vicinanze. Una comunitá virtuale, invece, é formata da utenti che stabiliscono un legame sui social media e iniziano ad interagire tra di loro. Il concetto di comunitá é stato ampiamente studiato in molti campi e, in particolare, nelle scienze sociali. Analizzare e studiare la formazione e l' evoluzione di comunitá é importante per molte ragioni. Per prima cosa, gli individui spesso formano gruppi in base alle loro affinitá e, quando si studiano gli individui, é interessante individuare questi gruppi. Infatti, si consideri, ad esempio, l' importanza per un rivenditore di un certo prodotto di trovare gruppi di persone con gusti simili al fine di raccomandare l'acquisto del prodotto. In secondo luogo, i gruppi forniscono una visione globale delle interazioni degli utenti, mentre una visione locale del comportamento individuale dei singoli é spesso rumorosa e ad hoc. Infine, alcuni comportamenti sono solo osservabili a livello di gruppo e non a livello individuale. Questo perché il comportamento dell'individuo puó fluttuare, mentre il comportamento collettivo di gruppo é piú robusto rispetto ai cambiamenti.

Gli oggetti costituenti un sistema in generale possono avere differenti livelli di interconnetti-

vitá. É importante tener conto di questi aspetti per comprendere meglio l'organizzazione di tali sistemi. É quindi necessario generalizzare la teoria delle reti tradizionali, che tiene conto solo di un tipo di relazione tra gli oggetti, a modelli e strumenti che permettano lo studio di sistemi multilivello in maniera adeguata. Nelle reti sociali, i collegamenti tra due oggetti possono essere classificati in base alla natura dell'interazione che essi rappresentano. Ridurre un sistema sociale ad una rete in cui gli attori sono connessi da un unico tipo di collegamento, spesso comporta una approssimazione della complessa realtá che si vuole rappresentare. É stata quindi riconosciuta dagli studiosi la necessitá di avere reti sociali che usano differenti tipi di collegamenti tra lo stesso insieme di individui. Queste reti sono denominate reti multilivello, in inglese multilayer, ma anche con termini alternativi come multidimensionali [19, 19, 13], multirelazionali [4, 11], multiplex [15, 8, 2].

L'obiettivo di questo documento é quello di fornire i concetti formali di base per la rappresentazione delle reti multilivello e la descrizione di una serie di misure, giá definite nel caso singolo layer, estese per le reti multilayer. Inoltre verranno descritti gli algoritmi proposti per la identificazione di comunitá in questo tipo di reti.

2.2 STRUTTURA DEL DOCUMENTO

Il documento é organizzato nel seguente modo. La prossima sezione introduce il concetto di rete multilayer. Nella Sezione 3 vengono introdotti i concetti base della teoria delle reti sociali multilayer, insieme ad alcune misure base. Nella Sezione 4 vengono descritti gli algoritmi per scoprire la struttura a comunitá di queste reti.

2.3 ACRONIMI E TERMINI CHIAVE

Reti multilayer, cammino, indici di centralitá, clustering.

3 MULTILAYER NETWORKS

Il concetto generale di rete multilivello é stato definito da Kivela et al. [12] come un grafo costituito da un insieme di nodi, un insieme di archi e un insieme di layer (livelli). Tali livelli possono o meno contenere alcuni nodi, collegati tra loro da archi all'interno un dato layer (intra-layer) o tra differenti layer (inter-layer). Tuttavia, per includere molteplici aspetti in una rete multilayer, vengono definiti un insieme di layer per vari tipi di interazione e un altro per rappresentare, ad esempio, timestamp temporali. Il termine "elementary layer" definisce un elemento di uno di questi insiemi, mentre il termine "layers" si riferisce a una combinazione di layer elementari da tutti gli aspetti. Ad esempio, un tipo di interazione rappresenta un layer elementare o un semplice timestamp, mentre una combinazione di un layer elementare con un timestamp determina un layer.

Una rete multilayer é dunque costituita da un numero d di aspetti e viene definita una sequenza $L = \{L_a\}_{a=1}^{d}$ di insiemi di layer elementari, uno per ogni aspetto. Un insieme di tutte le combinazioni di layer elementari $L_1 \times L_2 \times ... \times L_d$ costituisce un insieme di layer. In tal modo é possibile introdurre $V_M \subseteq V \times L_1 \times ... \times L_d$ come l'insieme delle combinazioni nodo-layer in cui un nodo é presente nel corrispondente layer, dove V é l'insieme complessivo dei nodi della rete. $E_M \subseteq V_M \times V_M$ definisce invece l'insieme degli archi della rete multilayer, ovvero l'insieme delle coppie di possibili combinazioni nodi-layer elementari.

Utilizzando il formalismo di cui sopra, una rete multilayer viene definita come una quadrupla $M = (V_M, E_M, V, L)$. Tale formalismo viene illustrato in Figura 1. L'insieme dei nodi é $V = \{1,2,3,4\}$, inoltre la rete possiede due "aspects" che hanno due corrispondenti elementary-layer $L_1 = \{A,B\}$ e $L_2 = \{X,Y\}$. Abbiamo quindi un totale di quattro differenti layer: $(A,X),(A,Y),(B,X),(B,Y)$. Ogni layer contiene un sottoinsieme di V. In tale esempio, l'insieme delle tuple node-layer é il seguente: $V_M = \{(1,A,X),(2,A,X),(3,A,X),(2,A,Y),(3,A,Y),(1,B,X),(3,B,X),(4,B,X),(1,B,Y)\} \subseteq V \times L_1 \times Ł_2$. I nodi possono essere collegati a coppia sia all'interno dello stesso layer (archi intra-layer, a linea continua) che tra i layer (archi inter-layer, tratteggiati).

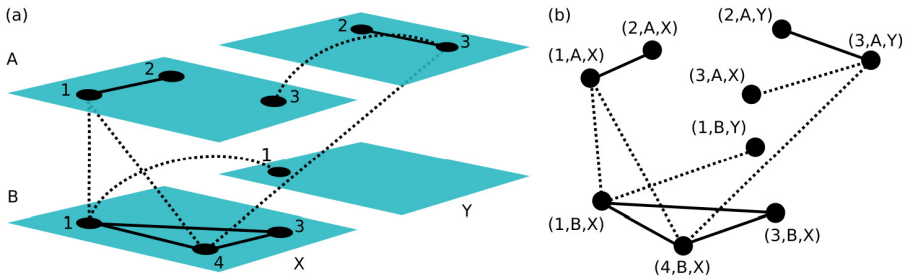

Figura 1: (a) Un esempio di rete multilayer, $M = (V_M, E_M, V, L)$. (b) Il grafo $G_M = (V_M, E_M)$ che corrisponde alla stessa rete multilayer.

Nel seguito vengono esaminate alcune misure utili per caratterizzare le reti multilayer.

La prima di esse é il **Node Degree** (grado del nodo) e il **Neighborhood** (vicinato). In modo del tutto generale, il grado di un nodo viene definito come il numero degli archi di qualsivoglia tipologia incidenti a quel nodo e il vicinato come l'insieme dei nodi raggiungibili da un nodo "focale" seguendo uno di quegli archi. Kivela et al. riprendono poi i concetti di "multidegree" (multi-grado) e di "multi-edge" (multi-arco) definiti da Bianconi in [3]. Un multi-edge viene introdotto utilizzando un vettore binario $m \in \{0,1\}^b$, che costituisce l'insieme delle coppie di nodi (u, v) tale che i nodi sono adiacenti (quando $m_\alpha = 1$) e non adiacenti (quando $m_\alpha = 0$) sul layer (livello) α.

Il **multidegree** k_u^m del nodo u é dato dal numero di multi-edge che il nodo u possiede con il vettore m. La **multistrength**, $s_{u,\alpha}^m$, é la somma dei pesi di quegli archi nella rete intra-layer di layer α.

Inoltre sono state introdotti i concetti di Walk (cammino), Path (percorso) e distanze. In particolare, un **Walk** e uno "step" (passo) sono dei concetti definiti tra coppie di tuple node-layer. Ció significa che esse considerano i collegamenti tra i differenti nodi all'interno dello stesso layer e anche tra i layer. Naturalmente questo ha senso solo nel caso in cui il passaggio da un layer a un altro venga considerato uno "step".

Un **Labeled Walk** é un walk associato a una sequenza di etichette dei layer. In tale contesto, la **Walk Length**, ovvero la lunghezza di un walk, viene definita in svariati modi. Essa viene indicata come un vettore che conta quanti step sono presenti in ogni layer. Un'altra definizione calcola la

walk length come la somma del numero totale di step intra-layer e inter-layer.

Infine i **Metapaths** sono considerati come sequenze di etichette associate ai nodi e agli archi che compongono il path. Nel caso in cui la rete multilayer venga aggregata e sia generata una versione "appiattita" della rete, é stata definita una **Distanza Aggregata** per ogni arco, al fine di calcolare gli "shortest path", ovvero i cammini piú brevi per raggiungere una destinazione.

L'**Interdipendenza** é invece il rapporto tra il numero di shortest path che attraversano piú di un layer e il numero complessivo di shortest path nella rete.

Riguardo il concetto di **Coefficiente di Clustering**, esso misura il grado di transitivitá di una rete monodimensionale. Un modo per definirlo consiste nel considerare la frazione di adiacenze esistenti rispetto a tutte le possibili adiacenze nel vicinato di un nodo. Tuttavia il concetto di vicinato e l'esistenza di connessioni a coppia posso essere intesi in molteplici modi nelle reti multilayer. Una seconda definizione del coefficiente di clustering é relativa al rapporto tra il numero di triangoli e le triple connesse. Il problema risiede stavolta nel concetto di "triangolo" che non ha una caratterizzazione univoca nelle reti multilayer. Infatti esistono molte possibili triple di nodi che contengono 3 nodi e 2 layers. Un terzo modo per definire il coefficiente di clustering consiste nel considerare i percorsi chiusi di lunghezza 3. Ma esistono molte definizioni di walk e path nelle reti multilayer. Alla luce di ció, sono state sviluppate delle generalizzazioni che utilizzano la prospettiva basata su densitá di un coefficiente di clustering tradizionale. É stata definita una coppia di coefficienti di clustering locali che partono da due modi alternativi di definire un vicinato di un nodo in una rete multilayer. Inoltre sono state introdotte definizioni alternative basate su differenti modi di definire un walk e un 3-ciclo in una rete multidimensionale.

Ancora, una **misura di centralitá** ha l'obiettivo di misurare l'importanza di un nodo, un arco o un sotto-grafo. Alcune generalizzazioni delle misure di centralitá a reti multilayer prevedono il calcolo dei valori di centralitá in corrispondenza delle tuple node-layer, mentre altre determinano solo valori di centralitá aggregati in corrispondenza dei nodi. La centralitá PageRank [18], basata sulla distribuzione stazionaria di un random walker su una rete monolayer, é stata generalizzata alle reti multilayer, in modo tale che il walker transiti all'interno di un layer o attraverso i layer.

Entrambi i nodi e i layer inoltre ricevono un rank. Un'altra generalizzazione di PageRank prevede un random walk in cui i passi intra-layer e inter-layer hanno differenti probabilitá. Un'altra misura di centralitá per reti multilayer prevede la costruzione di una matrice $nb \times nb$ basata sulle matrici di adiacenza intra-layer, su cui vengono calcolati gli autovettori. Infine, un modo per sviluppare nuove misure per le reti multilayer consiste nel confrontare le reti intra-layer di due o piú layer.

Ad esempio, il **Global Overlap** conta il numero di archi condivisi tra due layer, oppure si puó definire un indice di correlazione tra gli elementi della matrice di adiacenza dei due layer. Inoltre il **Grado di Molteplicitá** conta il numero delle coppie di nodi con archi di svariate tipologie tra di loro, diviso per il numero totale di coppie di nodi adiacenti. Un'altra definizione di grado di molteplicitá consiste nel considerare il numero dei layer in cui il nodo si trova.

4 AGORITMI PER L'IDENTIFICAZIONE DI COMUNITÁ IN RETI MULTILAYER

Sebbene le reti del mondo reale siano spesso multidimensionali, la ricerca si é focalizzata principalmente sulle reti monodimensionali. Molti approcci per individuare comunitá in reti con un solo tipo di collegamento tra due nodi sono stati proposti, e diverse rassegne sono state pubblicate [7, 10, 9]. Negli ultimi anni, tuttavia , l'interesse per reti complesse che presentano molteplici connessioni tra coppie di individui é in aumento, principalmente a causa della rapida crescita del social networking online. In realtá, le persone si connettono e interagiscono tra loro utilizzando una varietá di social media, e svolgono diverse attivitá che generano molteplici relazioni. Nel seguito vengono descritti alcuni degli approcci piú recenti.

Le proposte piú recenti per trovare gruppi in reti multidimensionali possono essere trovate in [19, 20]. In particolare, in [19] Tang et al. osservano che ci sono due approcci principali per gestire le reti multidimensionali. Il primo é una strategia naive che considera una rete multidimensionale come unidimensionale, utilizzando la rete di interazione media tra i nodi. Questa si ottiene calcolando la matrice di adiacenza

$$\overline{A} = \frac{1}{d}\sum_{i=1}^{d} A_i$$

come la somma media di tutte le d matrici di adiacenza $A_1, \ldots, A_d$, per ogni dimensione. $\overline{A}$ puó essere poi utilizzata da qualsiasi algoritmo noto di scoperta di comunitá. In particolare, gli autori applicano un metodo spettrale [5, 16], chiamato *Massimizzazione della Modularitá Media (AMM)* per ottimizzare il concetto di modularitá di Girvan e Newman [17]. L'altro approccio, chiamato *Massimizzazione della modularitá totale (TMM)*, consiste nell'ottimizzare la funzione obiettivo su tutte le dimensioni. Poiché Tang et al. usano la modularitá, si propongono di ottimizzare la modularitá totale

$$\overline{Q} = \frac{1}{d} \sum_{i=1}^{d} Q_i$$

Oltre a questi due approcci, propongono un nuovo metodo, denominato *Massimizzazione della Modularitá Principale (PMM)* che consiste di due fasi. In primo luogo, per ogni dimensione, le *caratteristiche strutturali* vengono estratte, quindi sono combinate per ottenere comunitá latenti. Tang et al. hanno esteso il loro approccio in [20] analizzando quattro diverse strategie per integrare caratteristiche strutturali. Uno dei principali inconvenienti degli approcci proposti é che il numero di comunitá deve essere dato come parametro di ingresso.

Li et al. [13], invece di trovare un partizionamento di una rete, affrontano il problema della costruzione di una comunitá per una rete multidimensionale, partendo da un nodo iniziale, e aggiungendo nodi confinanti alla comunitá seme purché essi siano simili. Zhang et al. [21] combinano reti sociali e contenuti generati dagli utenti per scoprire le comunitá di utenti che sono densamente connesse e, allo stesso tempo, condividono interessi comuni in contenuto. Comar et al. [6] hanno proposto un approccio per raggruppare reti multiple fattorizzando le loro matrici di adiacenza congiuntamente. In particolare, la matrice di adiacenza corrispondente ad un grafo viene scomposta in un prodotto di fattori latenti ed un metodo iterativo viene eseguito per minimizzare la funzione di distanza totale tra ciascuna matrice e il prodotto dei suoi fattori latenti. *MetaFac* é un approccio proposto da Lin et al. [14] per estrarre struttura di comunitá da dati sociali multidimensionali, rappresentati come tensori congiunti di dati, in cui ogni combinazione si realizza attraverso un multigrafo. L'approccio decompone tensori in matrici contemporaneamente, applicando la divergenza KL come misura di costo approssimazione. Il numero di scomposizioni generalmente non é noto in anticipo.

Un nuovo metodo, denominato *MultiGA* (*Multilayer Genetic Algorithm*), in grado di rilevare una struttura di comunitá condivisa in una rete multilayer é stato proposto in [1]. *MultiGA* adotta una rappresentazione genetica di individui che permette coevoluzione e cooperazione tra tutti i livelli della rete. Un individuo é composto da un numero di elementi pari al numero di strati. Ogni

elemento rappresenta una divisione dello strato corrispondente in comunitá, e si evolve contemporaneamente con tutti gli altri livelli 'imparando' da essi la loro struttura a comunitá attraverso l'ottimizzazione di una particolare funzione di fitness che combina l'informazione proveniente da ogni livello.

5 CONCLUSIONI

In questo documento é stato introdotto il concetto di rete multilayer, si sono descritte le proprietá che le caratterizzano ed é stata data una panoramica sugli approcci proposti recentemente in letteratura per l'identificazione di comunitá. Questi concetti sono basilari per comprendere le tecniche che verranno utilizzate per la definizione di servizi innovativi.

6 INDICE DELLE FIGURE

Elenco delle figure

Riferimenti bibliografici

[1] Alessia Amelio and Clara Pizzuti. A cooperative evolutionary approach to learn communities in multilayer networks, submitted, 2014.

[2] Federico Battiston, Vincenzo Nicosia, and Vito Latora. Metrics for the analysis of multiplex networks, August 2013.

[3] Ginestra Bianconi. Statistical mechanics of multiplex networks: Entropy and overlap. *Phys. Rev. E*, 87:062806, Jun 2013.

[4] Deng Cai, Zheng Shao, Xiaofei He, Xifeng Yan, and Jiawei Han. Community mining from multi-relational networks. In *9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005)*, pages 445–452, 2005.

[5] F. R. K. Chung. Spectral graph theory. *CBMS Regional Conference Series in Mathematics*, 92, 1997.

[6] Prakash Mandayam Comar, Pang-Ning Tan, and Anil K. Jain. A framework for joint community detection across multiple related networks. *Neurocomputing*, 76(1):93–104, 2012.

[7] L. Danon, J. Duch, A. Arenas, and A. Díaz-Guilera. Community structure identification. *Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science,World Scientific*, pages 93–113, 2007.

[8] M. De Domenico, A. Sole, S. Gómez, and A. Arenas. Random walks on multiplex networks. *arXiv:1306.0519*, 2013.

[9] S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

[10] S. Fortunato and C. Castellano. Community structure in graphs. *Encyclopedia of Complexity and Systems Science- Robert A. Meyers (Ed.)Springer*, pages 1141–1163, 2009.

[11] Andreas Harrer and Alona Schmidt. Blockmodelling and role analysis in multi-relational networks. *Social Netw. Analys. Mining*, 3(3):701–719, 2013.

[12] Mikko Kivelä, Alexandre Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer Networks, September 2013.

[13] X. Li, M.K. Ng, and Y. Ye. Multicomm: Finding community structure in multi-dimensional networks. *IEEE Transactions on Knowledge and Data Engineering*, in press, 2013.

[14] Yu-Ru Lin, Jimeng Sun, Hari Sundaram, Aisling Kelliher, Paul Castro, and Ravi B. Konuru. Community discovery via metagraph factorization. *TKDD*, 5(3):17, 2011.

[15] Peter J. Mucha, Thomas Richardson, Kevin Macon, Mason A. Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

[16] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3), 2006.

[17] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review*, E69:026113, 2004.

[18] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[19] Lei Tang, Xufei Wang, and Huan Liu. Uncoverning groups via heterogeneous interaction analysis. In *The Ninth IEEE International Conference on Data Mining (ICDM'09)*, pages 503–512, 2009.

[20] Lei Tang, Xufei Wang, and Huan Liu. Community detection via heterogeneous interaction analysis. *Data Mining and Knowledge Discovery*, 25(1):1–33, 2012.

[21] Zhongfeng Zhang, Qiudan Li, Daniel Zeng, and Heng Gao. User community discovery from multi-relational networks. *Decision Support Systems*, 54(2):870–879, 2013.