Analisi aggregata per ore

Michele Carignani, Alessandro Lenzi

2 marzo 2014

1 Generazione dei grafi orari

Per prima cosa i dati sono stati aggregati per ora. I dati originali del dataset Telecommunications - MI to MI sono nel formato:

```
timestamp \t SourceId \t DestId \t Stregth
```

e sono stati suddivisi in 24 file (uno per ogni ora) e aggregati, per cui ogni file contiene (al massimo[1]) un record per ogni nodo nel formato:

```
SourceId \t DestId:Strength [\t DestId:Strength]
```

A questo punto i pesi sugli archi (sopra chiamati `Strength`) sono stati riscalati rispetto alla somma dei valori della stella uscente di un nodo, ottenendo la probabilità di transire dal nodo i al nodo j, ovvero:

$$sumStrength_i = \sum_{j \in FS(i)} Strength_{ij}$$

$$probability_{ij} = \frac{Strength_{ij}}{sumStrength_i}$$

2 Ricerca delle componenti fortemente connesse

Per ricercare le componenti fortemente connesse (in seguito CFC) sono stati utilizzati sia diversiti tipi di taglio degli archi, sia diverse strategie di visita.

2.1 Tagli

Un taglio degli archi su un valore x significa utilizzare per la visita solo gli archi con peso (ovvero valore di probabilità) maggiore o uguale a x. Sono stati eseguiti test sui valori 0.001, 0.005, 0.007, 0.009, 0.01, 0.05, 0.07, 0.08, 0.09 e 0.1.

Vedremo i risultati dei tagli 0.005 e 0.05.

2.2 Strategie di visita

Le strategie di visita utilizzate sono 5 e impiegano i dati del dataset Telecommunications - SMS, Call, Internet - MI: a partire dai record del formato

```
SquareID \t Timestamp \t .. ChiamateInUscita ..
```

per ogni ora sono stati generati file con record

```
SquareID \t AggregatedCalls
```

[1] poichè certi nodi possono non avere chiamate in uscita in una certa fascia oraria.

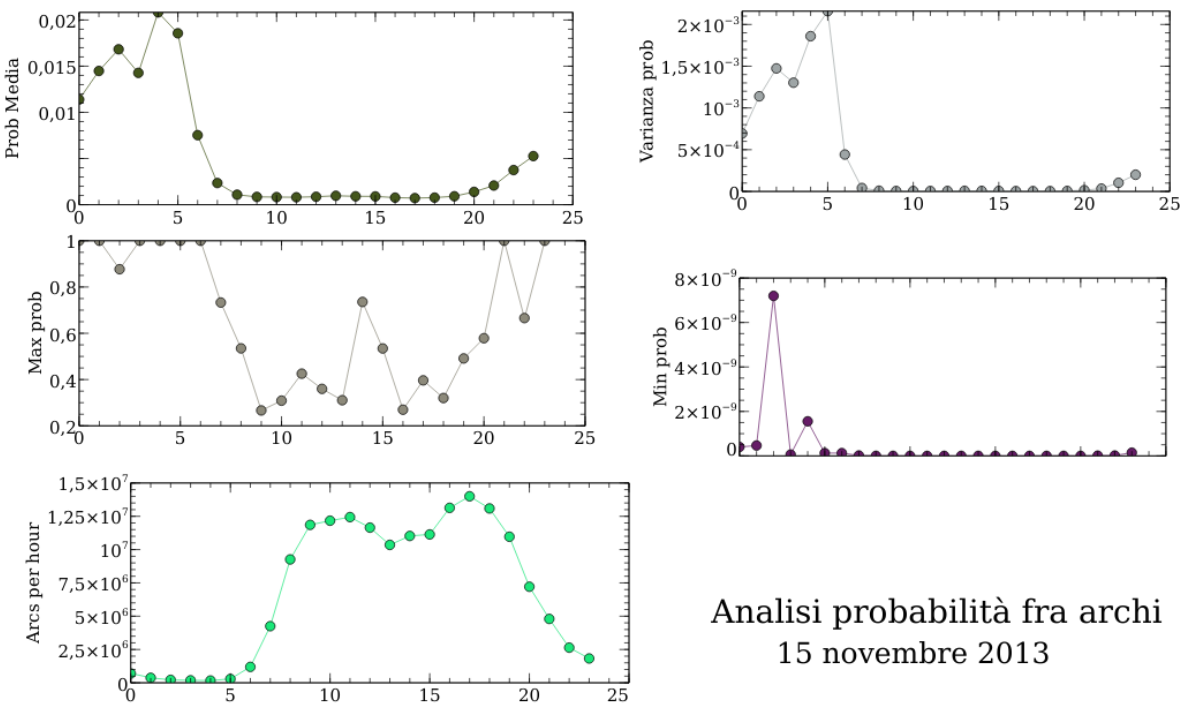

Figura 1: Statistiche sui pesi degli archi, 15 novembre. Sulle ascisse le fascie orarie, sulle ordinate le probabilità.

che permettono di capire quale sia il valore assoluto proporzionale a tutte le chiamate in uscita dallo square ID in una certa fascia oraria. In questo modo è possibile iniziare la visita del grafo non da nodi ordinati lessicograficamente ma in ordine (crescente o decrescente) di traffico in uscita. Le strategie inoltre si differenziano per il modo di ordinare la stella uscente da un nodo. Perciò le strategie definite sono:

- SCC1: visita il grafo in ordine crescente di traffico uscente, selezionando prima gli archi con probabilità maggiore;
- SCC2: visita i nodi per traffico decrescente e con archi selezionati per probabilità crescente;
- SCC3: visita i nodi per traffico decrescente e gli archi per probabilità crescente;
- SCC4: visita dei nodi per traffico crescente e archi per probabilità decrescente;
- Stable: Esegue la ricerca delle componenti fortemente connesse selezionando i nodi in ordine crescente di traffico telefonico **giornaliero** uscente e gli archi in ordine di probabilità decrescente.
- Stable con percentili: come sopra, ma tagliando il grafo utilizzando dei percentili.

3 Risultati

3.1 Statistiche

In fig. 1 sono mostrate le statistiche sui pesi degli archi come probabilità.

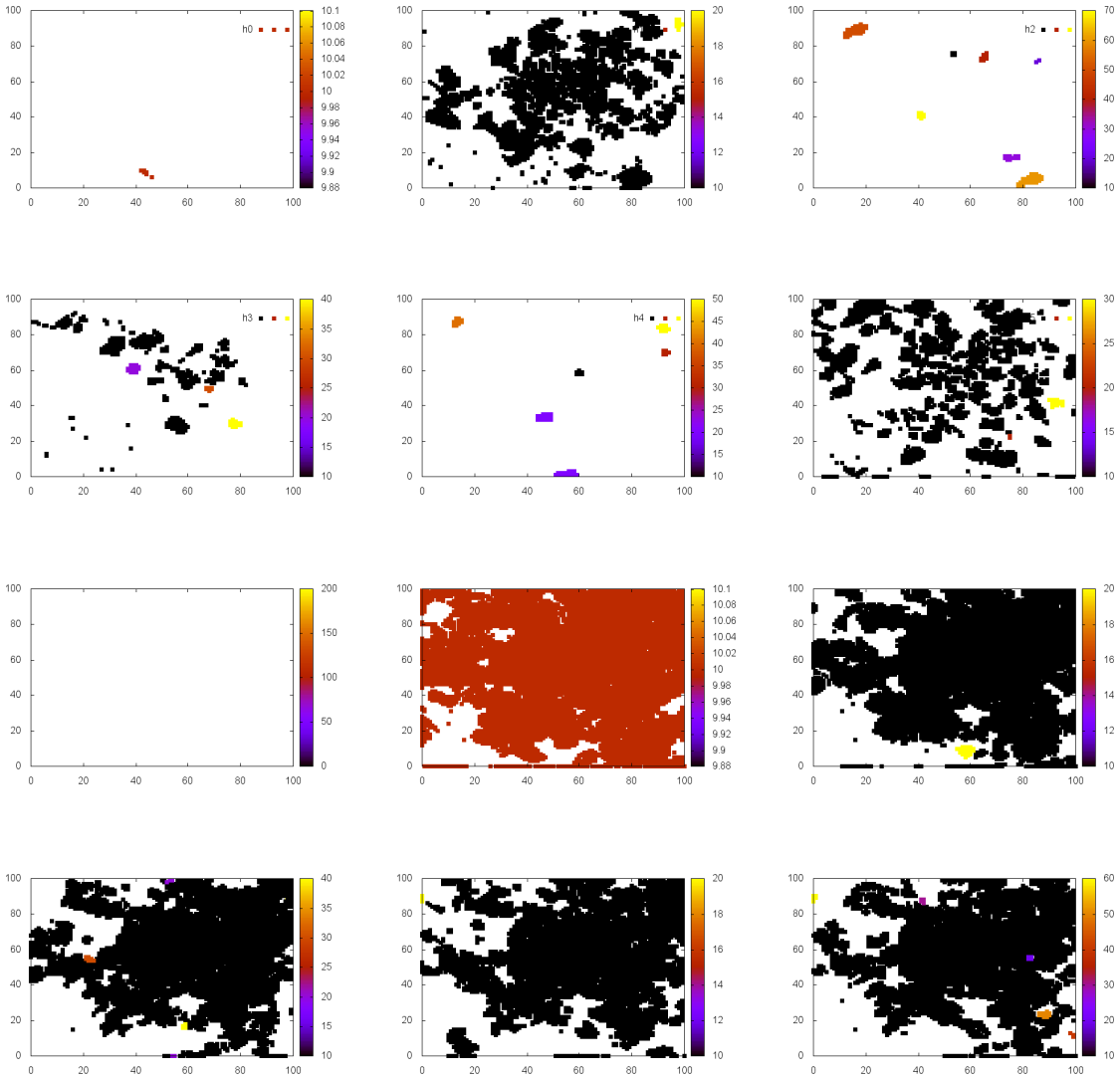

Figura 2: SCC1, Taglio 0.005, h 0-11. Da vedere da sinistra verso destra e dall'alto verso il basso

3.2 Componenti Fortemente Connesse

Le CFC sono state disegnate su mappe (utilizzando `gnuplot`) assegnando una funzione z alle celle di coordinate (x, y) così definita:

$$z(x,y) = \begin{cases} 0, & \text{se } (x,y) \text{ non appartiene ad alcuna CFC,} \\ 10k, & \text{se } (x,y) \text{ appartiene alla k-esima CFC trovata.} \end{cases}$$

3.2.1 SCC1, taglio 0.005

In questo caso, le componenti sono per la maggior parte delle ore estremamente ampie, come si può vedere in seguito: Si noti in 4 come la dimensione delle componenti può variare anche notevolmente a seconda del traffico presente in un'ora. Questo potrebbe essere dovuto al fatto che il taglio è estremamente basso, e pertanto agisce solamente su una parte minima degli archi. Nelle ore in cui il traffico è minore, anche le componenti hanno dimensioni notevolmente minori: probabilmente questo è dovuto al fatto che il numero di archi uscenti (fuori dall'orario lavorativo) è minore e con probabilità maggiore.

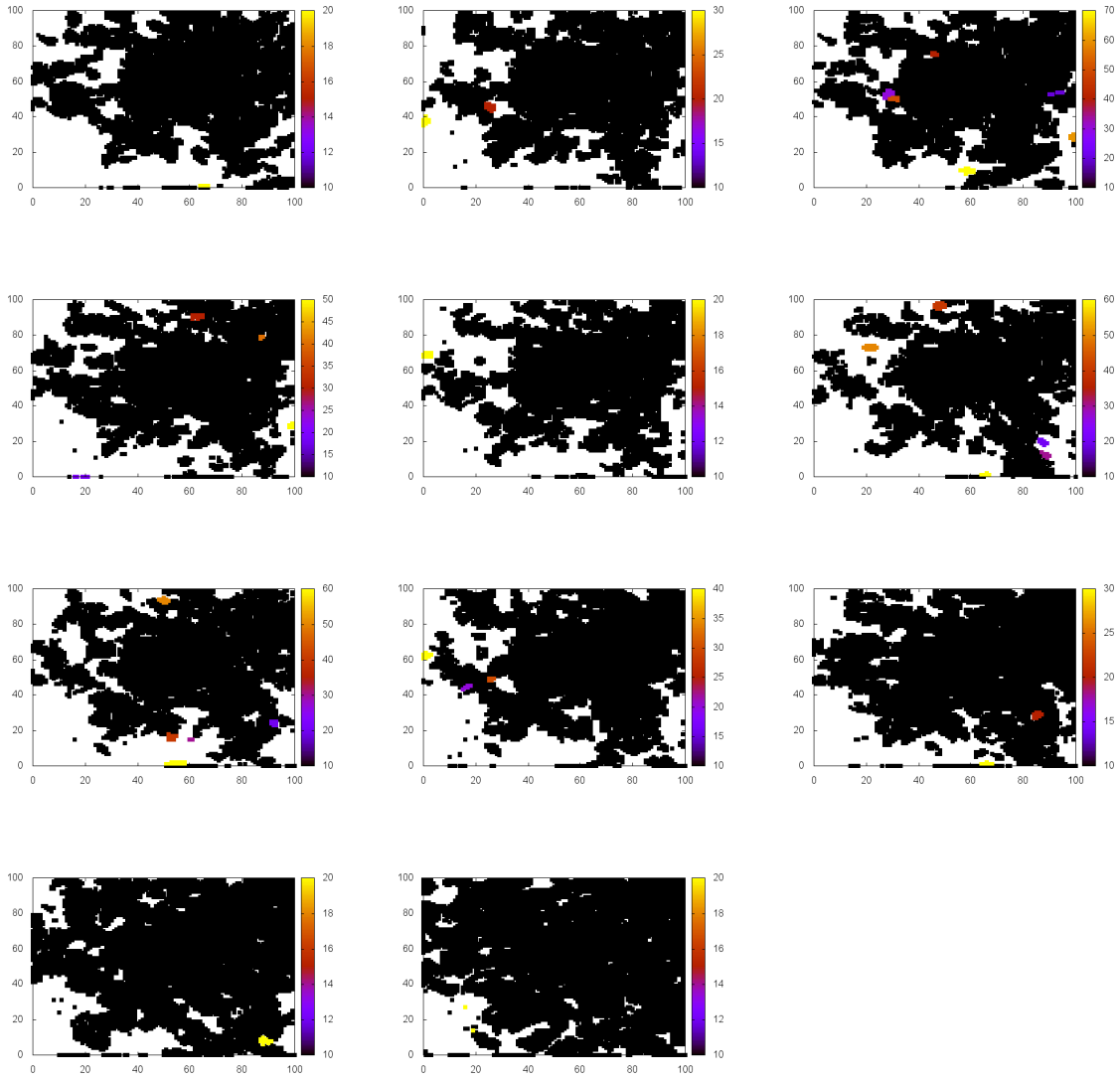

Figura 3: SCC1, Taglio 0.005, h 12-22; da vedersi da sinistra verso destra e dell'alto verso il basso.

```
0:      6,
1:      2972,11,
2:      4,2,13,10,38,40,9,
3:      462,19,7,16,
4:      7,41,7,14,14,
5:      2563,2,18,
6:
7:      8010,
8:      6605,29,
9:      5677,5,8,5,
10:     5638,4,
11:     5492,4,5,3,13,6,
12:     5729,4,
13:     5692,14,11,
14:     5818,5,13,5,8,6,12,
15:     5784,5,12,3,5,
16:     5362,11,
17:     5118,9,7,15,14,5,
18:     5623,8,2,11,11,14,
19:     5951,5,5,10,
20:     6676,11,6,
21:     7449,12,
22:     7971,2,
```

Figura 4: Nel listato, per ogni ora (alla sinistra), una lista delle dimensioni delle CFC trovate

3.3 SCC1, taglio 0.05

In questo caso, invece, le dimensioni delle componenti fortemente connesse diminuiscono notevolmente. Si noti, a conferma dell'ipotesi fatta in 3.2.1, come le dimensioni delle CFC calino drasticamente nelle ore di maggior traffico, mentre le ore che hanno un traffico inferiore sono caratterizzate da CFC più ampie. Si veda 5 Il risultato ottenuto, anche in questo caso non è stato considerato soddisfacente. Come esempio, mostriamo in 6 le componenti fortemente connesse trovate alle ore 10 e alle ore 11. Oltre alla loro dimensione - certamente non eccessiva - si può vedere che le CFC in 10 scompaiono in 11. Il dato confortante è che le CFC sono contigue geograficamente; ciò parebbe confermare l'ipotesi di una correlazione tra frequenza delle chiamate e vicinanza geografica.

```
0:      4,7,2,2,6,
1:      383,2,4,7,7,6,6,9,5,9,2,
2:      6,4,6,11,5,4,4,2,8,8,
3:      5,6,5,11,9,
4:      6,4,8,6,6,
5:      8,4,3,5,8,3,7,2,3,
6:      718,4,2,10,9,3,6,3,7,3,6,7,3,
7:      115,6,2,2,3,2,2,2,3,4,
8:      2,3,
9:
10:     2,2,
11:     2,
12:     3,
13:     2,
14:
15:     4,
16:
17:     2,
18:     4,5,5,
19:
20:
21:     2,2,2,2,4,5,2,5,
22:     248,2,2,2,2,4,2,6,6,4,2,2,2,2,9,
```

Figura 5: Da leggersi come in 4

```
==== 10.cfc
{7249,7250}
{7724,7824}

====11.cfc
{3659,3660}
```

Figura 6: Le CFC trovate alle ore 10 e alle ore 11 del 15 Novembre 2013