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.

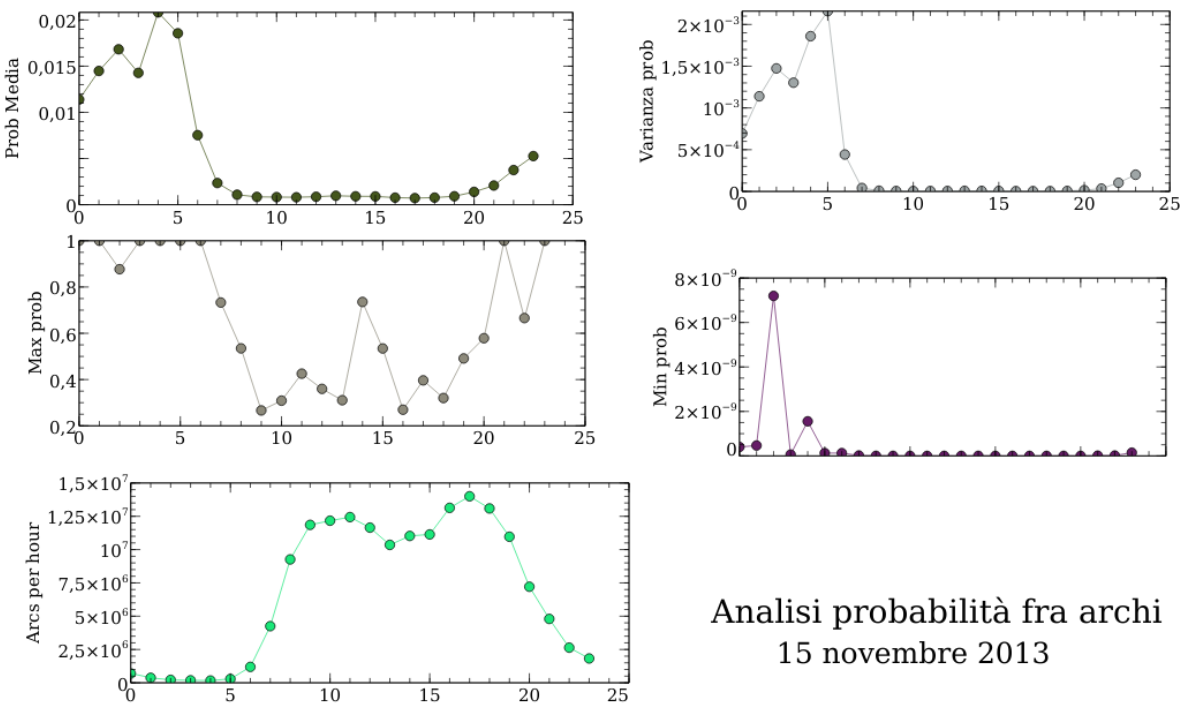

Figura 1: Statistiche sui pesi degli archi, 15 novembre. Sulle ascisse le fascie orarie, sulle ordinate le probabilità.

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 crescente;
- Stable: Esegue la ricerca delle SCC usando lo stesso ordine per tutti i file.

3 Risultati

3.1 Statistiche

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

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

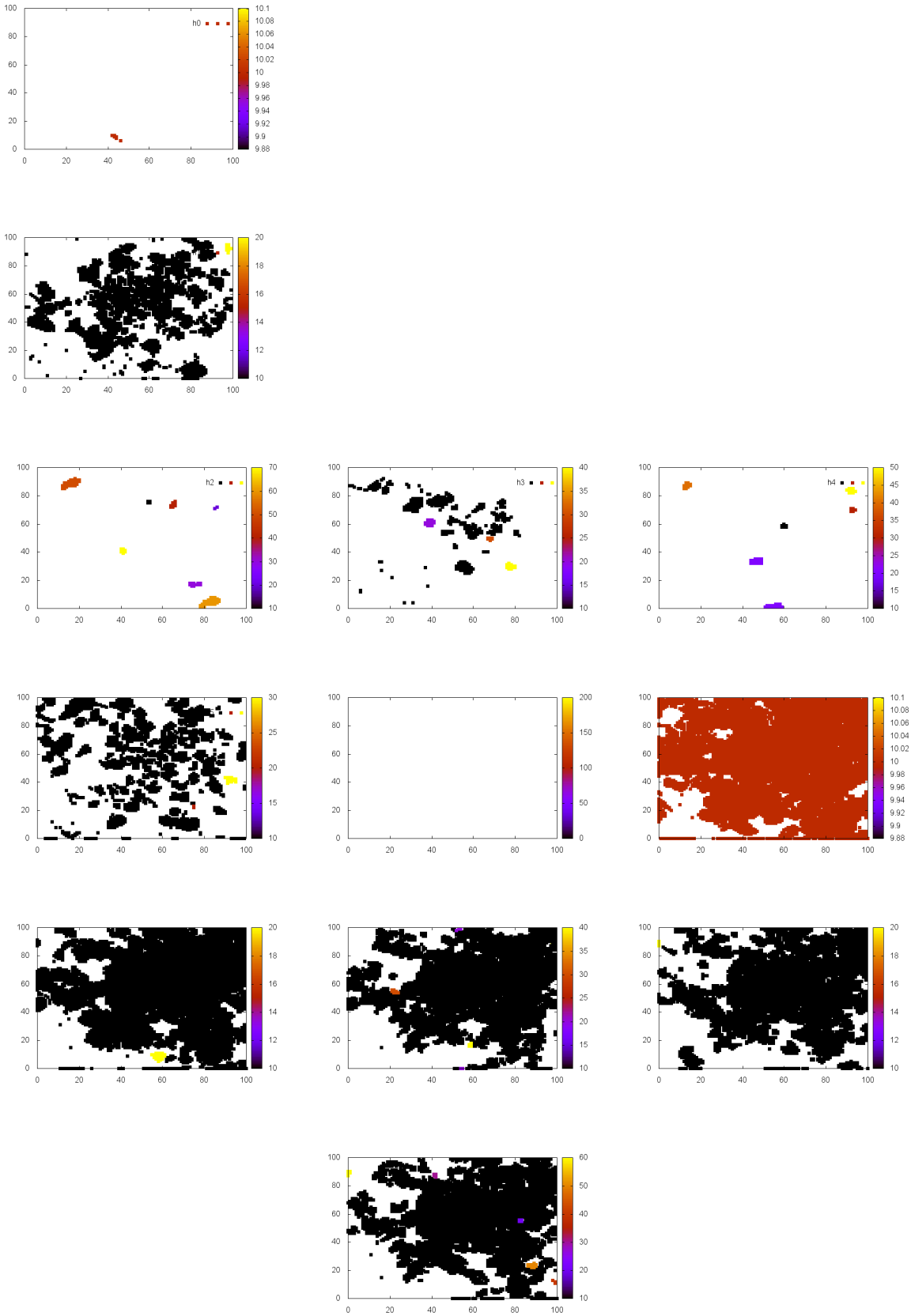

Figura 2: SCC1, Taglio 0.005, h 0-11

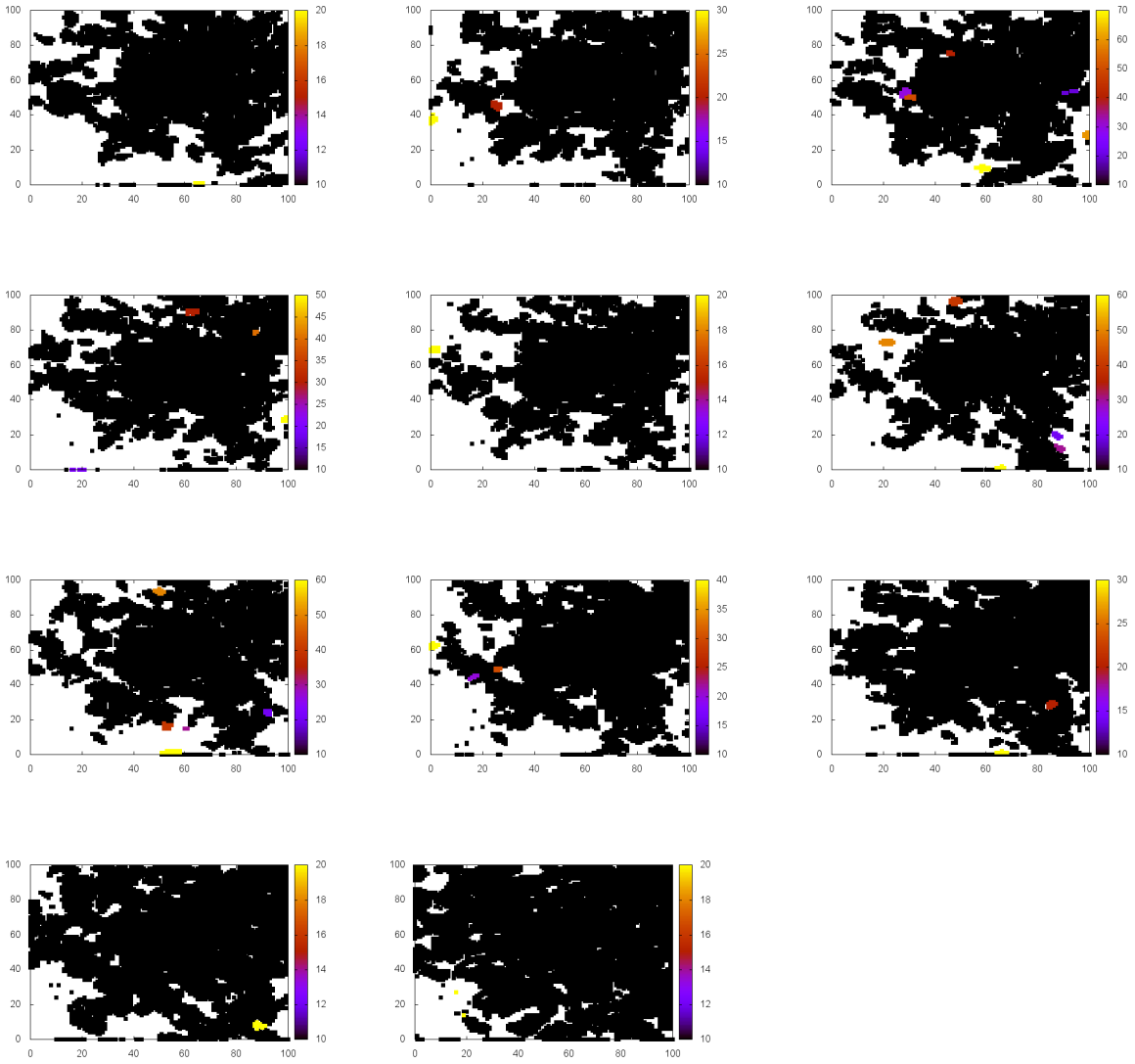

Figura 3: SCC1, Taglio 0.005, h 12-22