Analisi Su Omofilia e chiusure Relazionari

Relazione di: Andrea Marisio, studente dell’Insubria , matricola 718989

Università Degli Studi Dell’Insubria

Relazione Esame Di Analisi E Riconoscimento Delle Reti Sociali

Sommario

[Svolgimento del progetto 1](#_Toc431908875)

[Informazioni generali sui dati analizzati 2](#_Toc431908876)

[Informazioni di contatto 3](#_Toc431908877)

[Informazioni 3](#_Toc431908878)

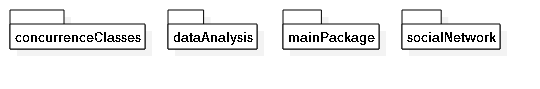
# Svolgimento del progetto

Il progetto è stato svolto in tre fasi:

* Sviluppo software per valutare i dati
* Analisi dei dati
* Utilizzo dell’API fornita da Gephi per la rappresentazione grafica dei dati

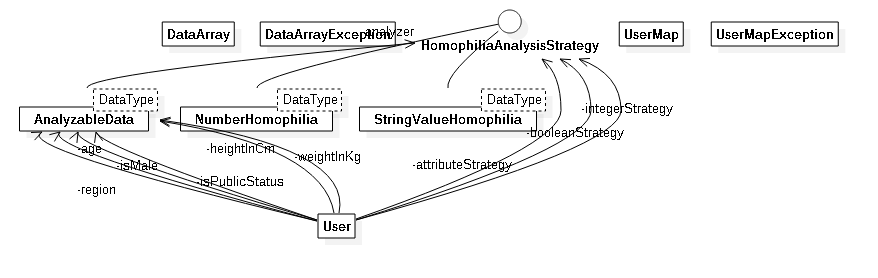
## Panoramica sul software

Il software di analisi sviluppato è stato sviluppato in Java, il progetto è diviso in 4 package

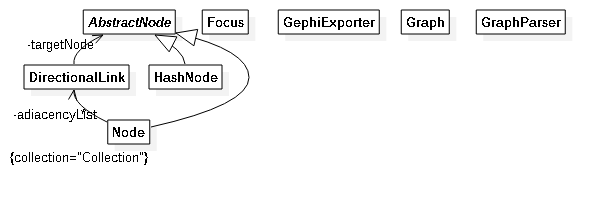


Dove dataAnalysis contiene classi necessarie per analizzare i dati ,come le strategie di analisi dei dati (le quali sono semplici strategie, per le stringhe l’uguaglianza , mentre per i numeri la vicinanza fino ad un massimo inserito come campo dell’oggetto). Per quanto riguarda User, questa classe ha una struttura che, per semplicità di implementazione, è legato alla struttura dei profili dei dati da analizzare, ma è possibile comunque generalizzarla.

La struttura del package è come segue :



Mentre per quanto riguarda network sono presenti dei componenti necessari per formare un grafo, e contengono metodi di analisi, come per esempio il calcolo della percentuale di triangoli chiusi, mentre altri come GephiExporter e GraphParser le quali hanno una utilità di interfaccia con altre strutture



## Informazioni generali sui dati analizzati

I dati analizzati provengono da <http://snap.stanford.edu/data/soc-pokec.html> il quale è uno snapshot della rete sociale slovacca Pokec. Questi dati sono composti da due file, uno relativo alle relazioni (sono relazioni direzionali) e uno riguardante le informazioni riguardanti i profili. Il primo file è un file “tab separated” (per intenderci, ID tab targetID) mentre per quanto riguarda il file relativo ai profili è un insieme di campi “tab separated” i quali a sua volta sono “comma separated” (es: 1 20 70kg,165cm Varese,Induno)

### I dati che ho analizzato

Vista la grande mole di dati, soprattutto i dati utente, per poter analizzare i dati richiedeva un tempo di calcolo elevato e per questa ragione ho provato a dividere le relazioni in base all’ID con il risultato che gli ID non sono stati assegnati in base alle relazioni, il 78% percento degli ID erano relazione fra ID molto differenti. Come alternativa ho provato a dividere gli utenti usando una ricerca in ampiezza valutando la divisione in base alla distanza così da avere il massimo numero di relazioni su numero di utenti, ma purtroppo la rete e molto densa il 65% delle relazioni è racchiudibile ad una distanza pari a due. Date queste informazioni ho scelto di dividere il grafo con una distanza pari a 3 e di analizzare i grafi generati più piccoli, così da avere un grafo abbastanza grande e veloce da analizzare

## Api fornita da gephi

Per rappresentare l’informazione avevo pensato di usare dei file CSV e poi utilizzare Gephi , ma i file CSV forniscono troppa poca informazione, quindi ho deciso di usare i file GEXF che sono dei file in XML , per la generazione di questi file, ho trovato una API fornita da Gephi, Gephi toolkit, che permette di fare tutto quello che fa il software Gephi direttamente nel sorgente che manipola i dati.

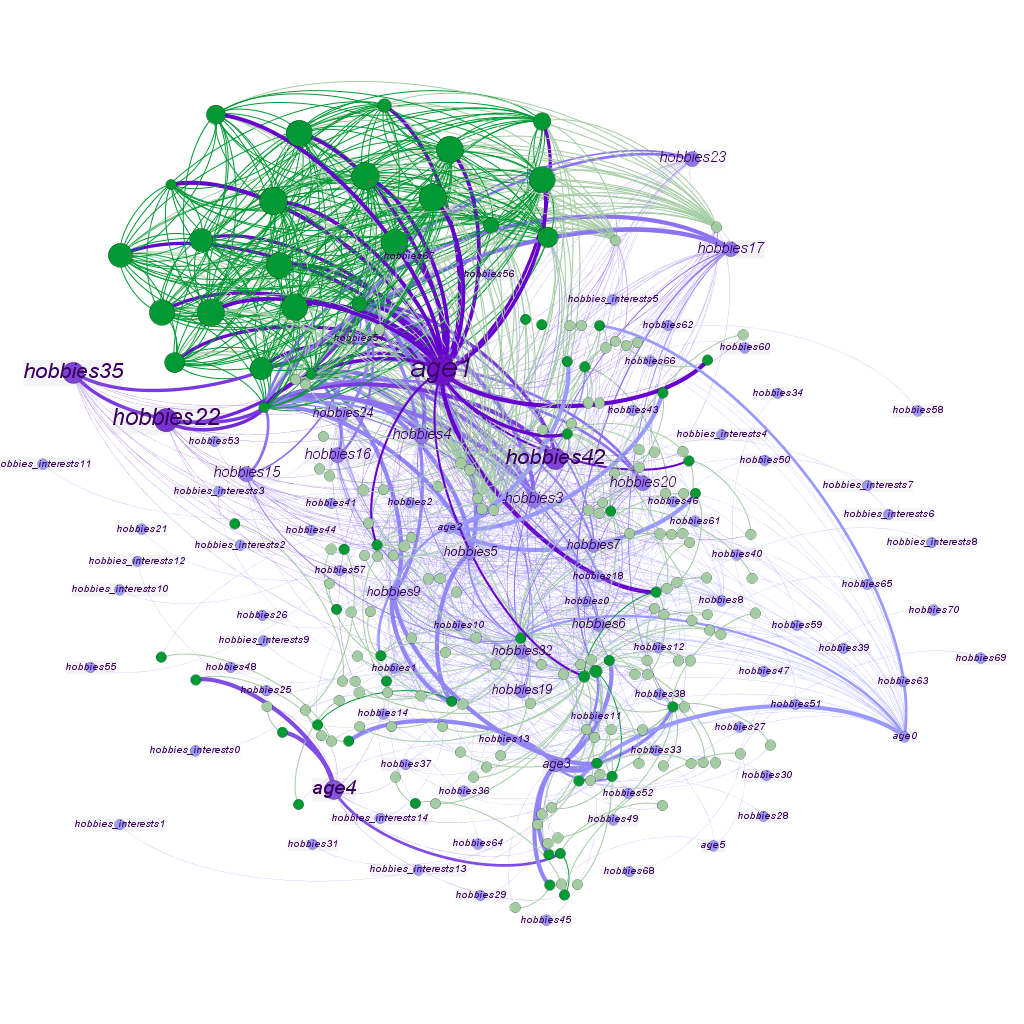
# Risultati Delle Analisi

La rete che vi sto per mostrare è estratta dalla rete pokec da circa 160 utenti con punti di aggregazione basati sull’età inserita e hobbies inseriti (rimappati per facilitarne la lettura, le età divise in decine e gli hobbie dalla scritta inserita).

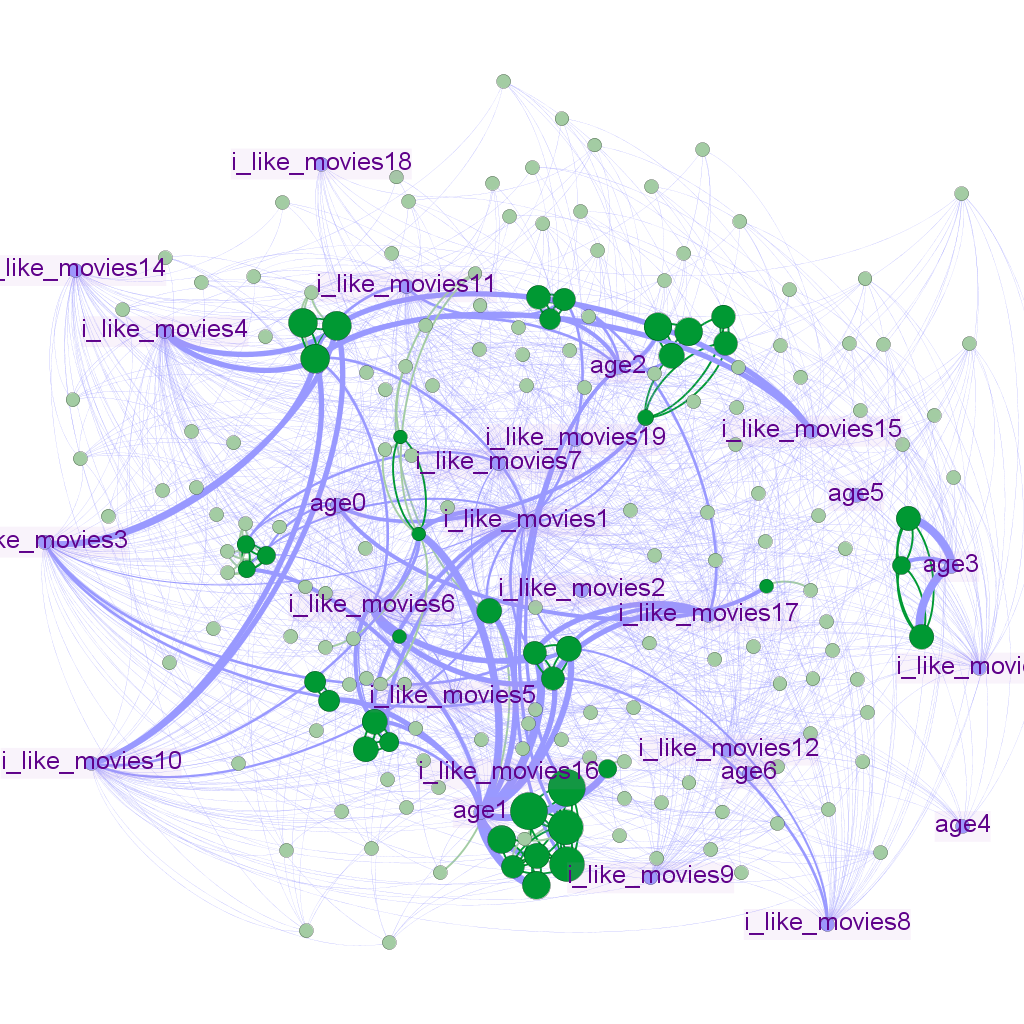
Il verde sta per gli utenti, più è scuro e più è alto il valore di omofilia media, mentre la dimensione è relativa alla media dei triangoli chiusi (chiusura triadrica), gli archi tra gli utenti hanno perso pari al valore di omofilia tra di loro.

Il viola, sta per i punti di interesse, la grandezza e il sua scurezza sono direttamente proporzionale alla chiusura focale, mentre il peso degli archi è la chiusura per appartenenza

Qui si nota la grande rilevanza dell’età (soprattutto age1), rispetto agli hobbies, ma si nota anche che la tendenza all’ omofila è più elevata, il numero di nodi chiari penso sia dovuto alla della sua inconsistente compilazione(es:age0, età inferiore a 10, spesso 0) o alla mancanza di molti campi.

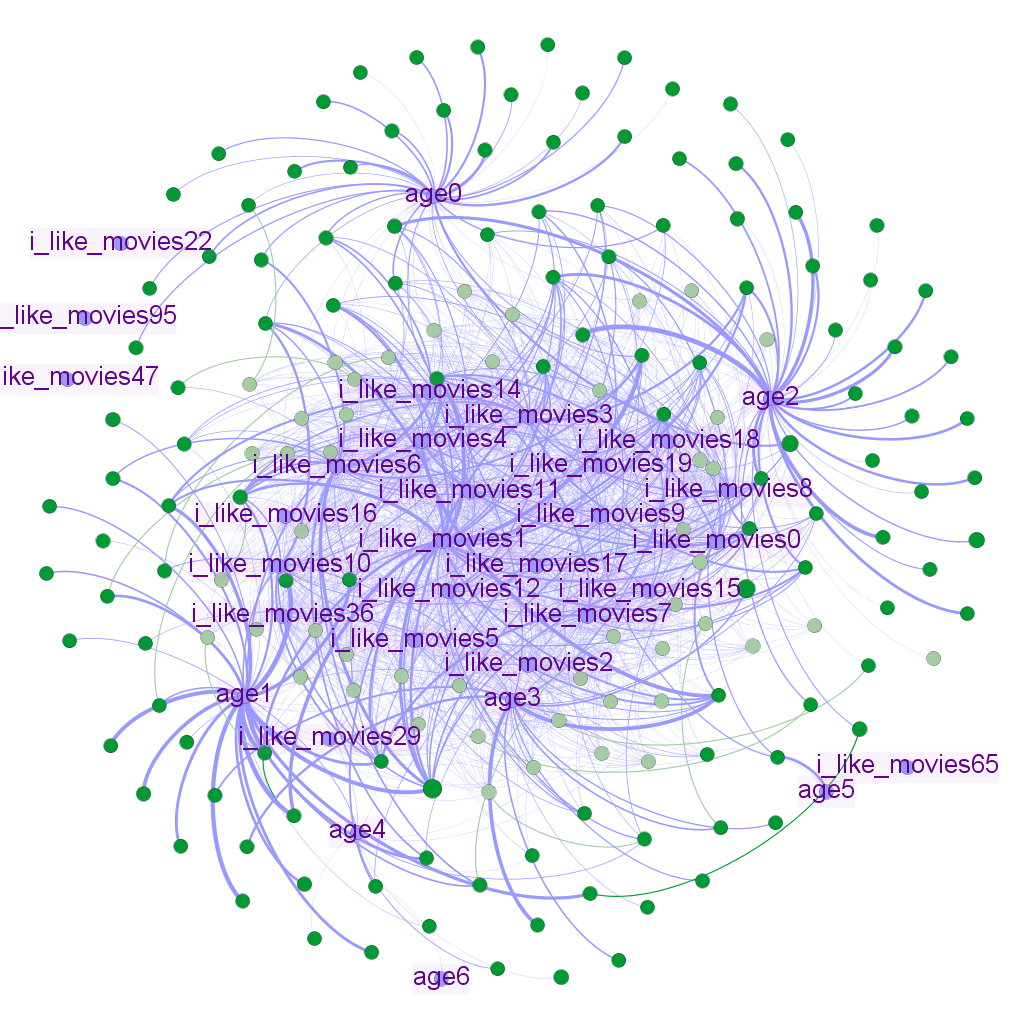


Se ci occupiamo di un grafo più grande, in questo caso un’estrazione da 15000 nodi circa, per poter mostrare i nodi, ho applicato dei filtri, in questo grafo ho preso i nodi più popolari (grado elevato) e si nota anche qui che il campo età ha un ruolo, la maggior parte dei nodi popolari con alto grado di omofilia media ha anche un elevato grado di chiusura dei triangoli

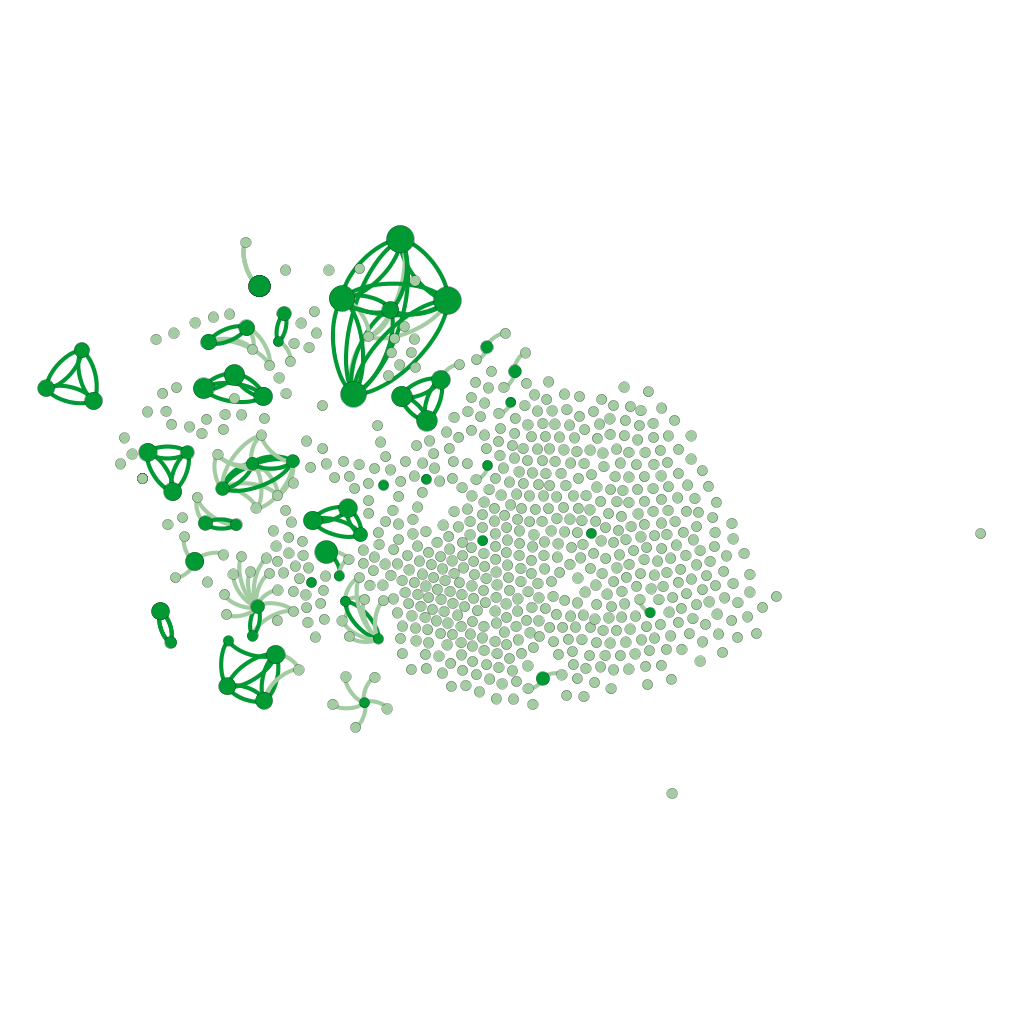


Prendendo invece i nodi con altissimo in-Degree e medio-alto out-Degree (non troppo alto per non escludere i punti focali, i quali hanno zero archi uscenti), si può notare che chi è molto “ricercato” non ha molte relazioni con altri “ricercati” ma nonostante tutto ha un livello di omofilia media rilevante e ha un basso livello di triangoli chiusi, quindi probabilmente è un ponte locale. Questi nodi sono di età inferiore a 30 anni (nel caso di age0, probabilmente hanno messo 0 sull’età), questi hanno amicizie con altri coetanei (si nota dallo spessore del l’arco con il focus, membership clousure)

Faccio notare che i legami tra questi nodi ci sono ma sono con un livello di omofilia basso se non nullo.

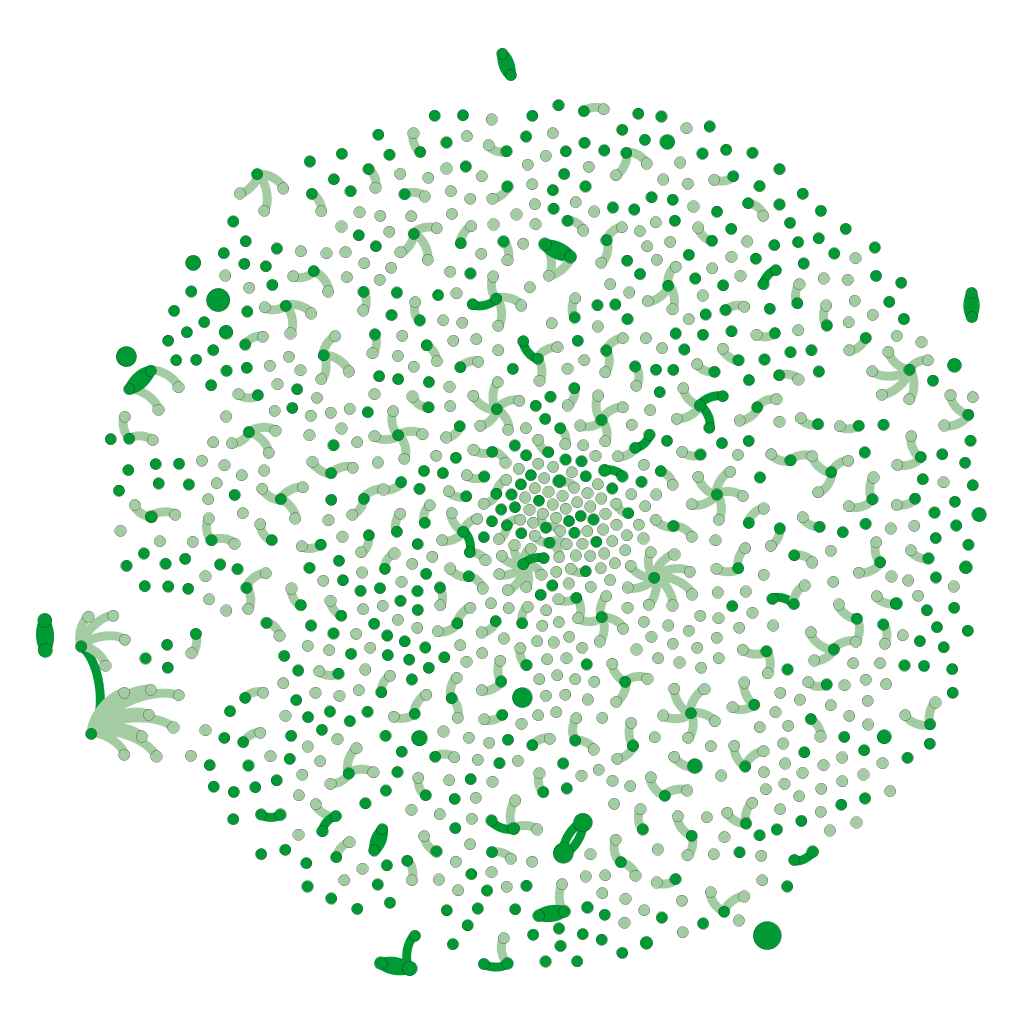


Se invece filtriamo la rete e prendiamo il i nodi con il top dei archi entranti, solo il 10% hanno una relazione con un altro con le stesse caratteristiche, ma quelli che hanno una relazione con uno di questi comporta una relazione di omofilia (il peso della relazione, lo spessore), un valore di omofilia medio e, alcuni, hanno un elevato numero di triangoli chiusi. Questi nodi, probabilmente, dispongono solo amici a per cui vale molto la relazione di omofilia, infatti le relazioni sono corrisposte.



Se invece prendiamo i top “puntatori” (nodi con un elevato out degree) è interessante notare che non ci sono nodi grandi, nodi con una percentuale elevata di triangoli chiusi ma, nonostante l’elevato numero di relazioni e il calcolo semplificato dell’omofilia, circa la metà possiede un livello medio di tendenza all’omofilia e

NodiTopOutDegreeDiscretoValoreDiOmofiliaPuntanoAPochePersoneUguali



# Informazioni di contatto

|  |  |  |
| --- | --- | --- |
| Andrea Marisio Studente Dell’Insubria |  |  |
| Sample person image |  |  | |  |  |
| **Matricola :** 718989  a.marisio@sudenti.uninsubria.it |  |  | |  |  |

# Informazioni

Università Degli Studi Dell’Insubria

Relazione Esame Di Analisi E Riconoscimento Delle Reti Sociali

http://informatica.dista.uninsubria.it/

