POLITECNICO DI MILANO
Facoltà di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria Informatica

Analisi della Costruzione Partecipativa di un Wiki con un'Applicazione a Wikipedia

Tesi di laurea di: Riccardo Tasso Matr: 708301

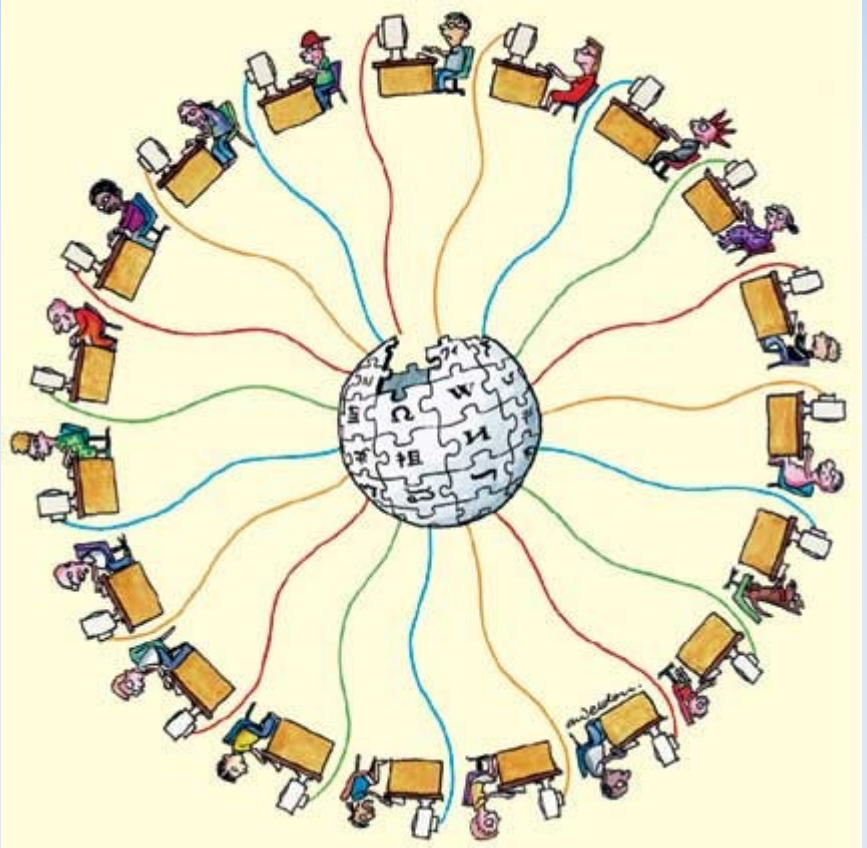
Relatore: Marco Colombetti
Correlatore: David Laniado

Anno accademico 2007-2008
Sessione di laurea del 20 aprile 2009

La tecnologia dei wiki

Nuova prospettiva del World Wide Web

- Il visitatore è anche *autore*
- Più autori *collaborano* alla costruzione di una conoscenza comune



Wikipedia

- Progetto di costruzione di un'enciclopedia
  - Online
  - Multilingua
  - A contenuto libero
- Il più grande esempio di wiki sul Web
  - Versione in italiano: 500.000 voci
  - Versione in inglese: 2.800.000 voci
- Enorme successo in breve tempo
  - Uno dei 10 siti Web più visitati al mondo
  - Tra i primi risultati dei motori di ricerca
  - Nascita di una comunità molto attiva

Il lato oscuro di Wikipedia

Quanto è **affidabile** il contenuto di Wikipedia?

Chi scrive su Wikipedia?

Quali sono i suoi **fini**?

Lavori di ricerca su Wikipedia

- Studi qualitativi
- Studi quantitativi
  - La crescita
  - Gli utenti
  - La misura dei contributi
  - La struttura dei collegamenti interni
  - I contenuti
  - La qualità

Obiettivi

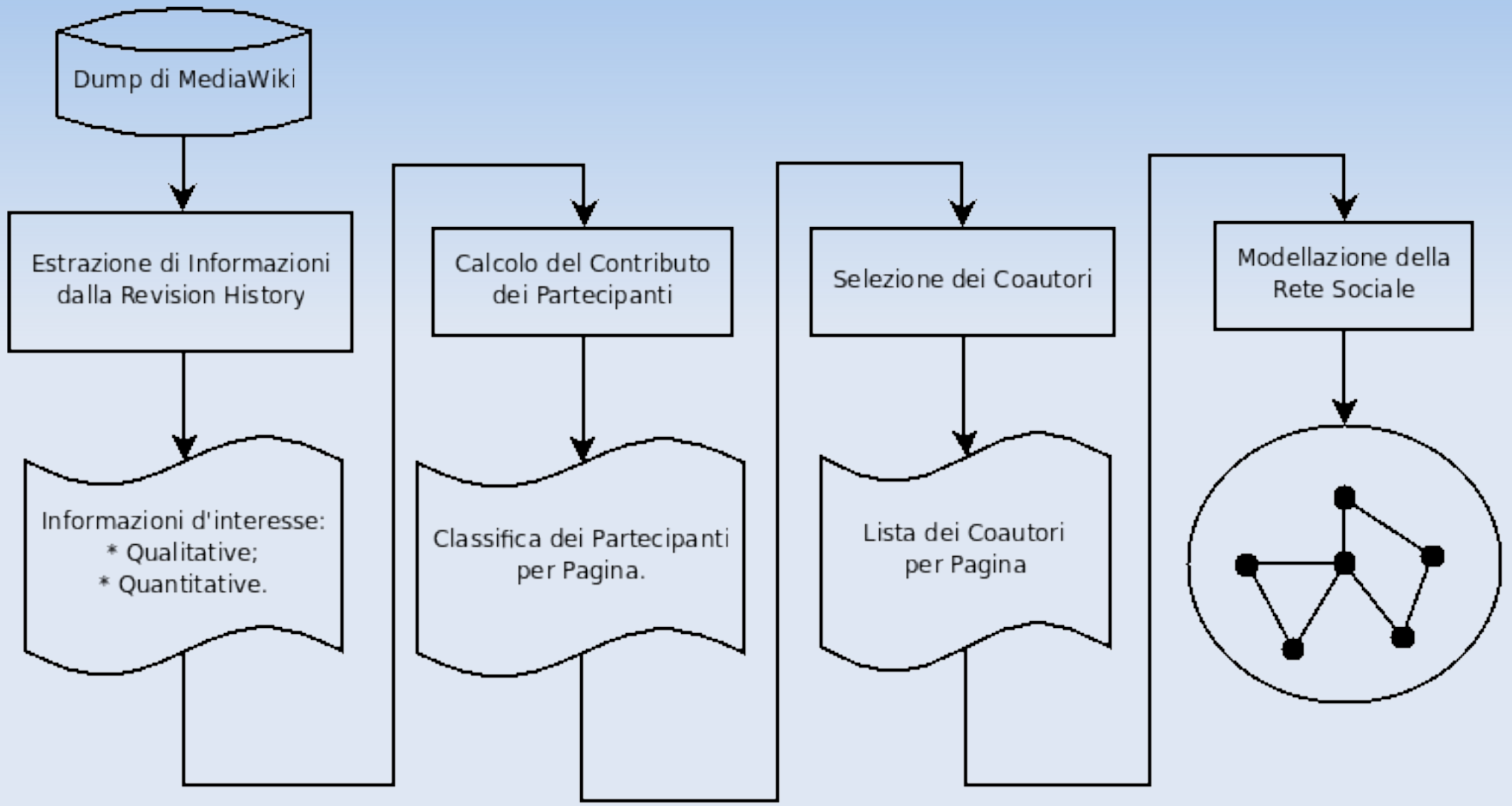
Progettare un metodo di analisi della comunità di un wiki:

- Automatico
- Generale
- Modulare

Studiare la comunità di Wikipedia:

- Per applicare il metodo a un caso reale
- Per capirne le caratteristiche generali
- Per capire come partecipano i suoi contributori

Il processo di analisi di un wiki



Estrazione di informazioni dalla Revision History

- Valutazione dell'operato di un utente a partire dai suoi interventi (*revision*)

- Informazioni ricavabili direttamente:
  - Timestamp
  - Numero di versione
  - Autore
- Informazioni non ricavabili direttamente:
  - Distanza tra due versioni: d(i, j)
  - Qualità di un intervento

La qualità di un intervento

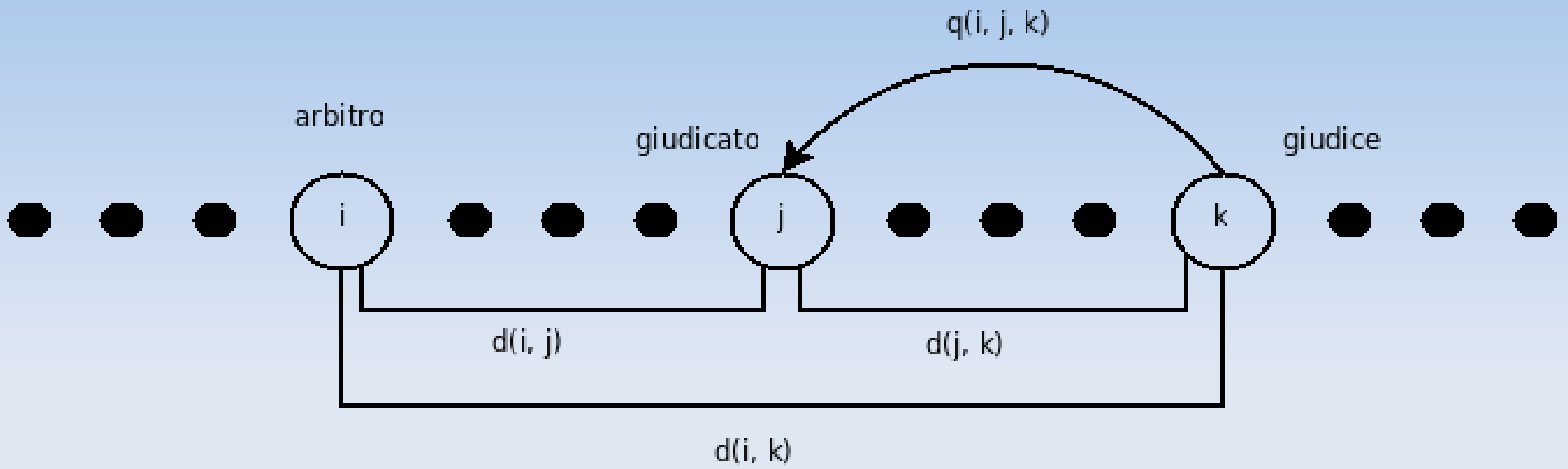
- La qualità di un intervento è stimabile
  - dalle modifiche che apportano gli interventi successivi
  - rispetto a quelle precedenti

Dante Alighieri è stato un poeta italiano.

Dante Alighieri è stato un poeta, scrittore e politico italiano.

Dante Alighieri è stato un poeta, **scrittore e politico** italiano. Egli nacque a Firenze nel 1265 ...

La qualità di un intervento



$$q(i,j,k)=\frac{d(i,k)-d(j,k)}{d(i,j)}$$

Calcolo del contributo dei partecipanti

- Livelli di granularità:
  - Globale
  - Singola pagina
- Metriche:
  - Conteggio degli interventi (edit count)
  - Longevità dell'intervento (edit longevity)
  - *Longevità dell'intervento valutata rispetto alla sua versione più simile*
- Problemi:
  - Come contare il contributo degli utenti anonimi?

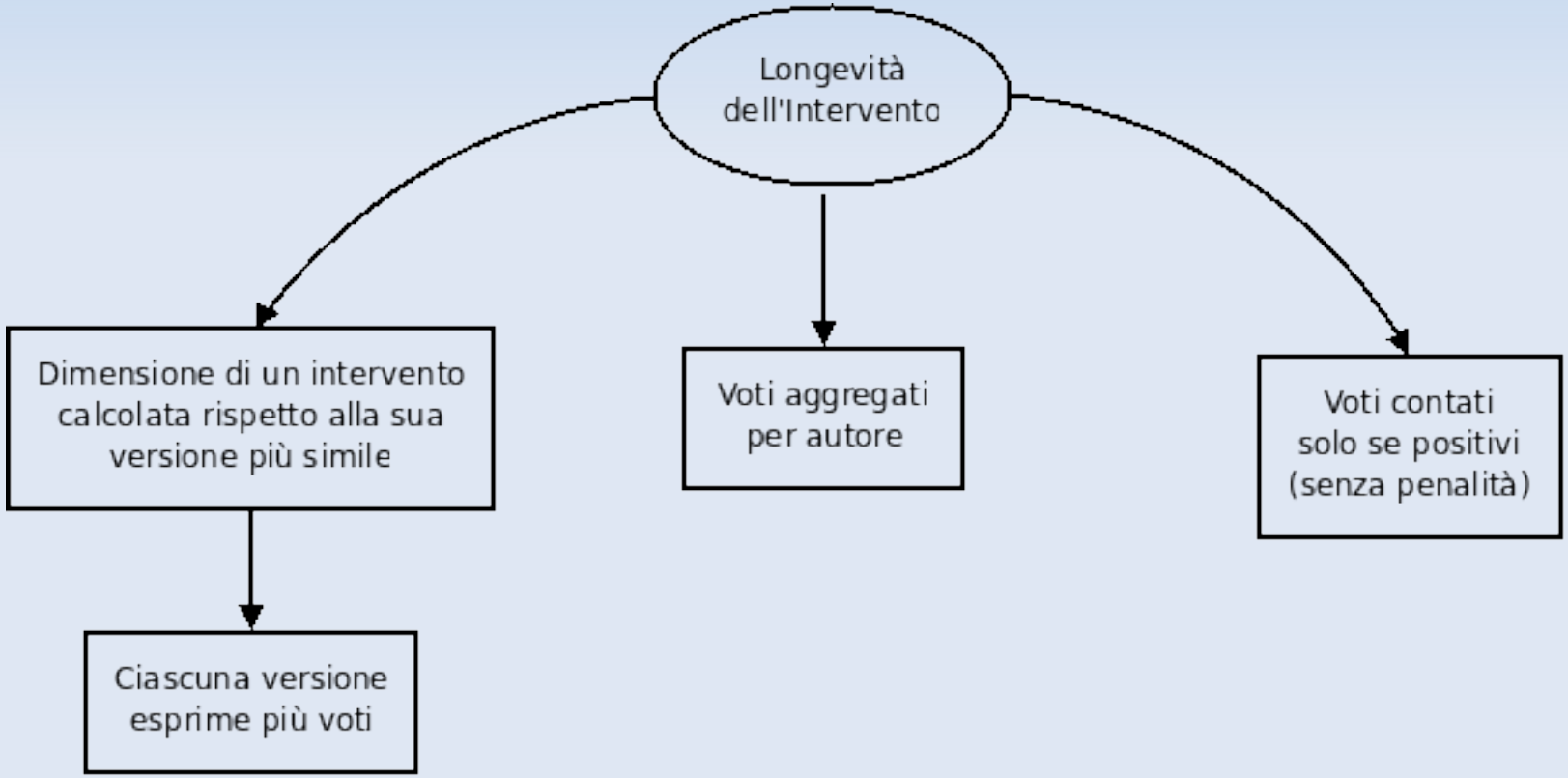
Calcolo del contributo dei partecipanti

- Conteggio degli interventi:
  - Semplice
  - Attualmente usato su Wikipedia
  - Poco preciso

- Longevità dell'intervento:
  - Tiene conto della **dimensione** di un intervento: distanza dalla versione precedente
  - Tiene conto della **qualità** di un intervento: valor medio dei voti ricevuti dalle 10 versioni successive rispetto alle 10 versioni precedenti

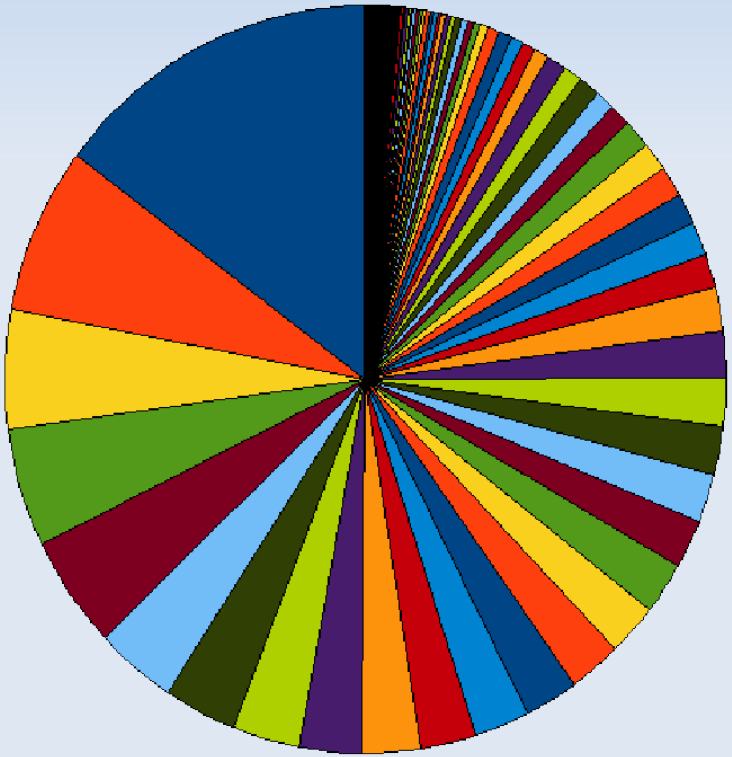
Calcolo del contributo dei partecipanti

- Longevità dell'intervento valutato rispetto alla sua versione più simile

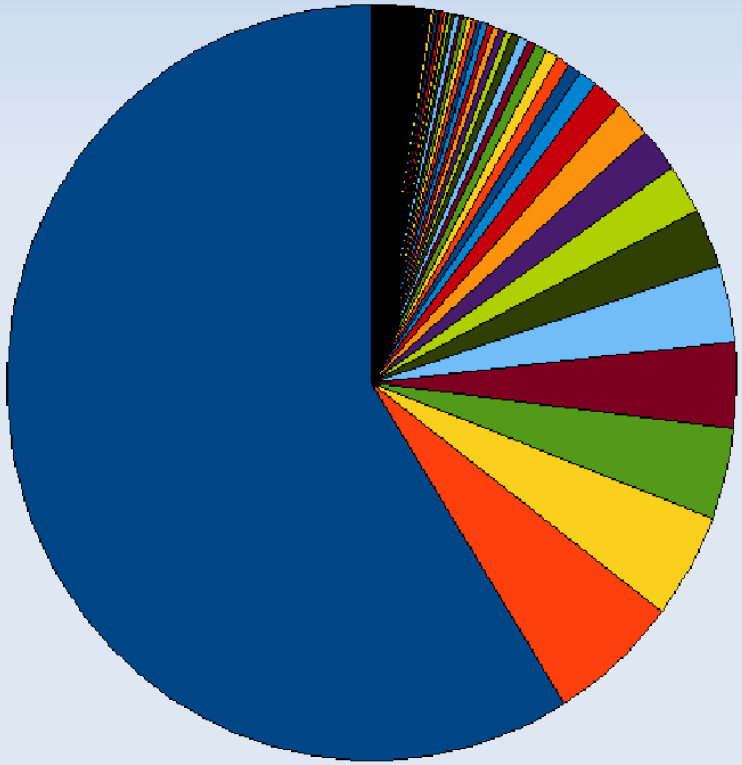


Selezione dei coautori

Pagine apparentemente molto simili possono essere state costruite in modo completamente diverso



Pagina: HIV
Numero di Edit: 1282
Numero di Utenti: 415



Pagina: C++
Numero di Edit: 1225
Numero di Utenti: 434

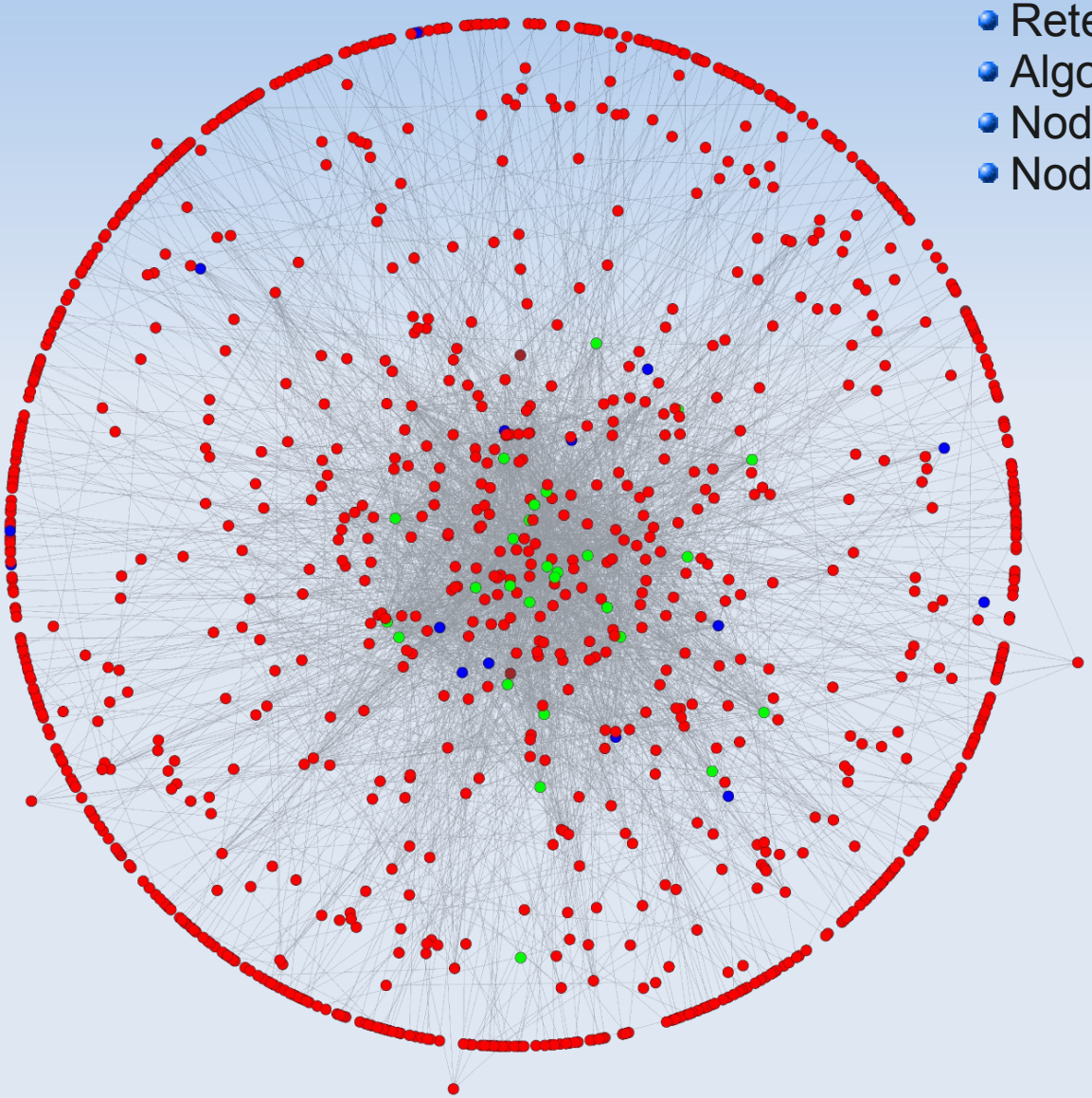
Selezione dei coautori

- Scegliere, in modo automatico, gli utenti che hanno contribuito maggiormente alla costruzione di una pagina
- Idee dell'algoritmo:
  - Esaminare gli utenti ordinati per contributo (decrescente)
  - Vincolo 1: selezionare il più piccolo insieme che raggiunge una certa percentuale di contributo (50%)
  - Vincolo 2: selezionare un utente solo se ha contribuito oltre una certa quantità (10)

Modellazione di una Rete Sociale

- Social Network Analysis
  - “Insieme di metodi per l'analisi di strutture sociali che consentono un'investigazione precisa degli aspetti relazionali di queste strutture” (Scott, 2000)
- Perché una Rete Sociale di Wikipedia?
  - Per studiare le caratteristiche della comunità nel suo insieme
  - Per studiare i ruoli degli individui all'interno della comunità
- Idea: ricondursi alle reti di coautori
  - Di articoli scientifici
  - Di software open source

Rete Sociale della versione italiana di Wikipedia (2005)



- Rete dei coautori di Wikipedia al 13.12.2005
- Algoritmo di visualizzazione: Spring Layout
- Nodi verdi: Amministratori
- Nodi blu: Bot

Misure effettuate sulle Reti Sociali

- Reti studiate:
  - Wikipedia in italiano al 13.12.2005
  - Wikipedia in italiano al 22.05.2007
  - Wikipedia in italiano al 17.03.2008
  - Wikipedia in inglese al 06.02.2007

- Misure relative alla costruzione della rete:
  - Numero di articoli esaminati
  - Articoli con almeno un autore
  - Articoli con più di un autore
  - Numero di autori
  - Articoli per autore
  - Autori per articolo

Misure effettuate sulle Reti Sociali

- Proprietà macroscopiche della rete:
  - Numero di autori (nodi)
  - Collaboratori per autori (grado medio)
  - Studio delle componenti connesse
  - Coefficiente di clustering
  - Distanza media tra due nodi
  - Diametro
- Studio delle Sociometric Star:
  - Degree Centrality
  - Closeness Centrality
  - Betweenness Centrality
  - Eigenvector Centrality

Risultati Sperimentali

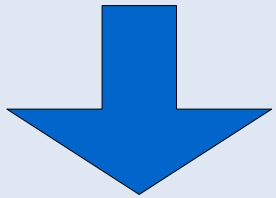
- Calcolo dei contributi a livello globale:
  - Il *conteggio degli interventi* è una metrica poco precisa e da particolare risalto ai Bot
  - La metrica di *longevità di un intervento* favorisce i manutentori del wiki (a causa dei revert)
  - La metrica di *longevità di un intervento valutata rispetto alla sua versione più simile* trova i contributori in modo più preciso (ma spesso non è molto diversa dalla precedente)
- Processo di selezione dei coautori:
  - Pagine apparentemente simili possono essere costruite in modo molto diverso tra loro
  - Insiemi di pagine dello stesso tipo (Featured) possono avere caratteristiche comuni molto diverse da quelle totali

Risultati Sperimentali

- Costruzione della Rete Sociale:
  - Confronto tra le versioni in italiano e in inglese di Wikipedia:
    - la prima è molto più piccola
    - in proporzione la comunità inglese è più piccola rispetto a quella italiana
  - Confronto tra versioni del medesimo wiki in istanti differenti:
    - la crescita della comunità è confermata
    - gli individui più centrali sono rimasti gli stessi
  - Le metriche di calcolo del contributo proposte non mostrano particolari differenze tra le reti costruite a livello macroscopico
  - Le misure di centralità studiate mettono in evidenza quasi sempre gli stessi autori → i ruoli all'interno della comunità sono poco distinti

Risultati Sperimentali

- Costruzione della Rete Sociale:
  - La misura di Eigenvector Centrality mostra però significative differenze rispetto alle altre → gli utenti centrali tendono a non collegarsi tra di loro
  - Vale il modello di *preferential attachment* → nuovi autori tendono a collegarsi a quelli con un maggior numero di collaboratori
  - Caratteristiche in comune tra le diverse versioni analizzate:
    - Basso coefficiente di clustering
    - bassa distanza media tra nodi
    - piccolo diametro



Lo scopo delle personalità più importanti di Wikipedia pare essere quello di non lasciare alcuna delle sue aree senza il loro intervento

Conclusioni

- Definizione di un metodo per l'analisi della comunità di un wiki:
  - Automatico
  - Generale
  - Modulare
    - Estrazione dei dati
    - Calcolo dei contributi (confronto tra diverse metriche)
    - Selezione dei coautori
    - Modellazione di una Rete Sociale
- Ciascun sottoproblema:
  - È stato studiato dal punto di vista formale / teorico
  - È stato affrontato progettando opportuni moduli software in grado di implementare le considerazioni teoriche
  - È stato applicato a quattro differenti versioni di Wikipedia
- I risultati hanno:
  - Dimostrato la fattibilità dell'analisi
  - Mostrato interessanti caratteristiche di Wikipedia

Sviluppi Futuri

- Costruzione di una Rete Sociale solo per certi sottoinsiemi di pagine
- Utilizzo delle informazioni contenute nel peso degli archi della rete
- Tecniche di clustering e community detection
- Studio dell'evoluzione della rete nel tempo
- Costruzione di una rete bipartita
  - Due tipologie di nodi: autori e pagine
- Applicazione del metodo ad altri wiki