Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea Magistrale in Scienze di Internet

Fidarsi di Wikipedia.
Attacchi vandalici e resilienza dei contenuti: un modello di analisi ed alcune evidenze empiriche

RELATORE:
Chiar.mo Prof.
Marco Ruffino

Presentata da:
Simone Dezaiacomo

Sessione I
Anno Accademico 2013/2014

More information (particularly thanks to the Internet) causes more confidence and illusions of knowledge while degrading predictability. At non time in the history of humankind we have lived in so much ignorance (easily measured in terms of forecast errors) coupled with so much intellectual hubris.

- Nassim Taleb

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Lo scopo dello studio è comprendere i fenomeni alla base della fiducia degli utenti verso l'enciclopedia online Wikipedia. Per farlo è necessario prima di tutto comprendere e modellizzare l'organizzazione della struttura dei processi socio-produttivi sottostanti alla produzione del contenuto di Wikipedia, procedendo quindi nel verificare empiricamente e descrivere le capacità di autocorrezione della stessa. Oltre a quelli utilizzati in questo studio, saranno anche descritti gli approcci e i risultati trattati in letteratura, riportando i principali studi che nel corso degli anni hanno affrontato questi argomenti, sebbene mantenendoli indipendenti.

Per comprendere la struttura della community degli editor di Wikipedia, si è ipotizzata l'esistenza di un modello di tipo core-periphery. Per studiare il modello sono state eseguite delle analisi su dati derivanti dalla versione italiana di Wikipedia[1], su un intervallo temporale completo, potenzialmente dal primo articolo scritto sull'enciclopedia; per limiti computazionali tuttavia non è stato possibile utilizzare l'intero dump di Wikipedia in lingua italiana, ma sono state selezionate solo alcune pagine appartenenti a differenti categorie ricavando così un campione eterogeneo il più possibile rappresentativo dell'universo di riferimento. Si è scelto di focalizzare lo studio sulla versione italiana di Wikipedia, sia per limiti nella capacità di elaborazione, sia per dare al lavoro un focus più specifico. Le analisi sui dati sono basate sui dump messi a disposizione dalla Wikimedia Foundation[2], nello specifico

[1] Wikipedia, L'enciclopedia libera, http://it.wikipedia.org

[2] http://wikimediafoundation.org

i dati utilizzati provengono dai dump XML del 8 Maggio 2014[3].

Le informazioni estratte dall'analisi di queste strutture rappresentano le basi utilizzate per la selezione delle pagine oggetto dell'iniezione degli errori, costituendo un metodo per stimare le diverse probabilità di autocorrezione per ciascuna pagina. Per quanto riguarda le capacità di resilienza di Wikipedia, i risultati sono ricavati utilizzando un approccio empirico che consiste nell'inserimento di errori all'interno del campione di pagine sotto specifici vincoli metodologici per poi valutare in quanto tempo e con quali modalità questi errori vengono corretti. Sono state effettuate specifiche analisi per la scelta delle tipologie di errori e delle variabili da considerare nell'inserimento di questi, ovvero la tipologia delle pagine e la metodologia di iniezione. Questa analisi ha portato alla definizione di 2 esperimenti tra loro distinti, i cui risultati portano ad interessanti conclusioni sia visti separatamente che combinati tra loro. Sulla base dei risultati di questi esperimenti è stato possibile discutere sulle capacità di autocorrezione del sistema, elemento chiave nello studio delle dinamiche della fiducia verso Wikipedia.

[3]http://dumps.wikimedia.org/itwiki/

Capitolo 1

Caratteristiche strutturali e dinamiche di Wikipedia

1.1 Wikipedia come oggetto tecnologico, sociale e di conoscenza

1.1.1 Storia di Wikipedia

> "My dream is that someday this encyclopedia will be available for just the cost of printing to schoolhouses across the world, including '3rd world' countries that won't be able to afford widespread internet access for years. How many African villages can afford a set of Britannicas? I suppose not many."
> *Primo messaggio di Jimmy Wales alla mailing list di Nupedia, 2000*

Wikipedia è un'enciclopedia libera online, probabilmente la più grande istanza di software Wiki. Il sito attuale deriva da un progetto ad esso precedente, Nupedia, con il quale condivideva l'obiettivo di creare un'enciclopedia gratuita online ma dal quale si differenzia nella metodologia di produzione dei contenuti, in quanto nel progetto originale questi sarebbero dovuti essere

generati tramite rigorosi processi di "expert review" sui contenuti scritti sotto licenza GNU Free Documentation Licence. Il progetto Nupedia, lanciato in Marzo del 2000 da Jimmy Wales e i collaboratori Sanger e Shell[41], si presentò subito come ambizioso, con lo scopo di diventare "the world's largest Encyclopedia". Per raggiungere questo scopo una delle maggiori difficoltà era quella di cercare collaboratori "esperti quasi in tutto", come riportato in un articolo di PC World del Marzo 2000[1] che descriveva e pubblicizzava il progetto Nupedia: "*The site's managers are seeking contributors and editors with expertise in, well, almost anything. The contributors will provide the diverse content, which will be offered free of charge to both consumers and businesses. Anyone is welcome to peruse Nupedia, and any other Web site may post Nupedia's content on its own. They need only to credit Nupedia as the source.*"

Il progetto ebbe un grande successo grazie anche all'infrastruttura tecnologica sui cui si appoggiava, poichè l'azienda di Wales (Bomis Inc.) aveva grandi esperienze e risorse nella progettazione e promozione di siti web ad alto traffico. Nel primo anno, i membri di Nupedia erano 602, di cui circa il 30% possedeva un Ph.D. A gennaio 2001 vi erano 2000 persone nella mailing list di Nupedia.

Nonostante l'iniziale successo, Nupedia si trovò presto in difficoltà. Gli sforzi necessari al recupero di esperti certificati e affidabili portava a rendere ancora più complesso il processo editoriale di Nupedia e la crescente necessità di pubblicare nuovi articoli e l'aumento dell'interesse nella collaborazione non erano rispecchiati dal processo editoriale di Nupedia. Fu quindi proposta e attuata l'introduzione dei software Wiki per dare la possibilità agli editori (anche non certificati) di collaborare sugli articoli prima di farli entrare nel processo di peer review vero e proprio e quindi di discutere sui testi candidati a diventare articoli ufficiali dell'enciclopedia[1]. Il portale, per necessità tecniche dell'epoca, non fu integrato nel sito di Nupedia stesso e nel Gennaio 2001 fu lanciato il sito parallelo Wikipedia.com. Con questa tecnologia, in

[1] L.E. Gouthro, PC World, 10 Marzo 2000

grado di fornire informazioni ad un costo relativamente basso e di permettere a collaboratori distanti di interagire con semplicità, Wikipedia cresce in fretta e attira addirittura più autori del progetto principale, che viene chiuso l'anno successivo

Nel corso degli anni Wikipedia si espande e vengono lanciate delle versioni anche in altre lingue oltre alla originale in lingua inglese. Nel 2005 si arriva a 195 lingue, di cui 60 con più di 1000 articoli e 21 con più di 10000. Ad oggi le statistiche mostrano che per l'intero progetto Wikipedia vi sono oltre 31 milioni di articoli [2] e 2 milioni di Wikipediani[3]. Per quanto riguarda i singoli progetti, il progetto solitamente preso come riferimento è l'originale Wikipedia inglese, con oltre 4,5 milioni di articoli e 1 milione di utenti attivi. Per quanto riguarda Wikipedia in lingua italiana gli articoli sono oltre 1 milione e circa 50000 gli utenti [4].

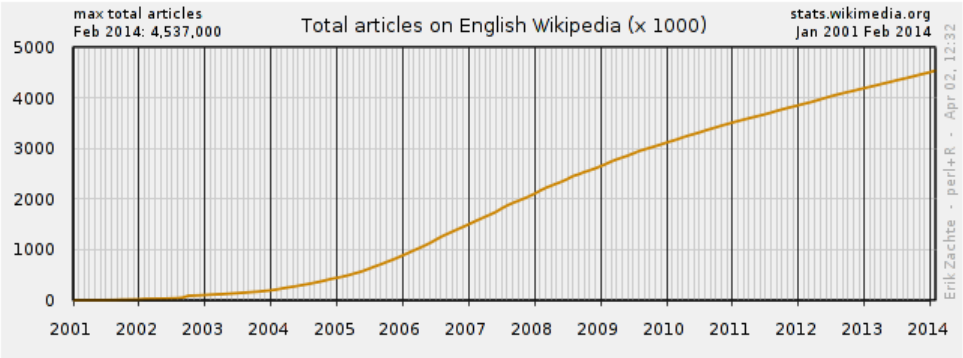


Figura 1.1: Totale articoli Wikipedia (Eng)

[2]articoli contenenti almeno un link interno

[3]utenti che hanno effettuato almeno 10 modifiche

[4]Statistiche ufficiali Wikimedia, http://stats.wikimedia.org/

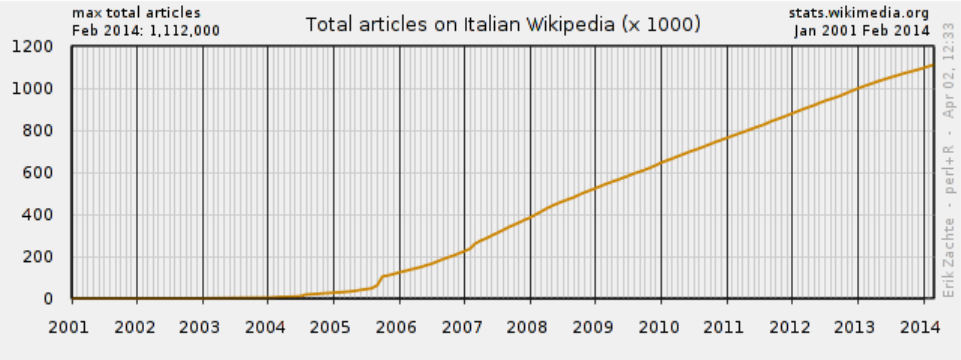


Figura 1.2: Totale articoli Wikipedia (Ita)

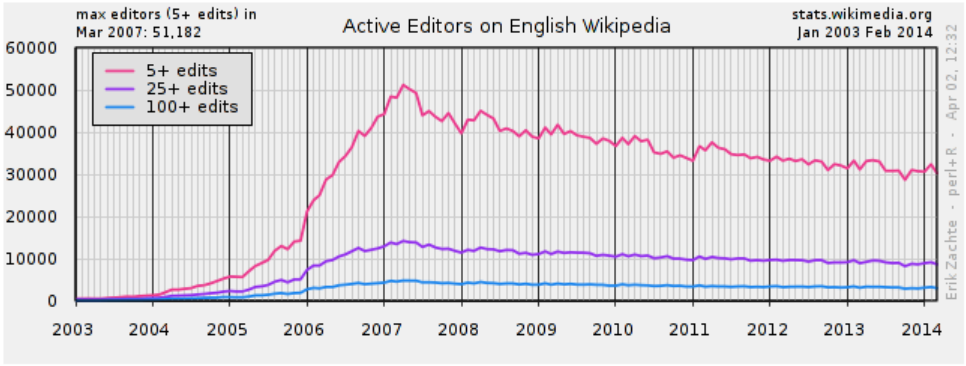


Figura 1.3: Totale editor attivi (Eng)

1.1.2 Chi scrive Wikipedia

Con l'introduzione dei software Wiki è stata invertita l'asimmetria dello schema produttivo del contenuto sul Web, spostandolo dal client al server. Di conseguenza, una pagina che prima poteva essere letta da tutti ma modificabile solo dal proprietario, ora diventa una pagina modificabile da chiunque attraverso il browser, inserendo il markup adeguato che sarà interpretato e tradotto dal software lato server che genererà la pagina HTML corrispondente. In media ci sono 10 modifiche al minuto nella Wikipedia in lingua italiana[5]. Il numero totale di autori per pagina è un valore difficile da mi-

[5]Statistiche ufficiali WikiMedia: http://stats.wikimedia.org/EN/SummaryIT.htm

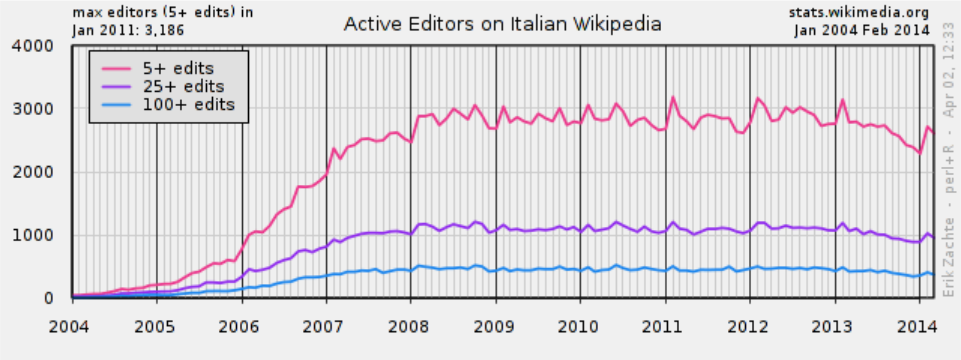


Figura 1.4: Totale editor attivi (Ita)

surare in quanto un singolo autore potrebbe possedere account multipli e Wikipedia permette la modifica in *modalità anonima* (viene traccia l'indirizzo IP ma questo non è un valore con caratteristiche tali da poter essere associato ad uno specifico utente). Le misurazioni ufficiali di Wikistats si basano sugli *active Wikipedians* intesi come quegli utenti che hanno effettuato almeno 5 modifiche nell'ultimo mese.

Ma Wikipedia non è semplicemente un'enciclopedia online, che permette a tutti di contribuire, il sito è in realtà l'artefatto più visibile di una comunità attiva ed emergente. Wikipedia è prima di tutto una comunità e l'enciclopedia è l'immagine e risultato delle *discussioni* di questa comunità; è la realizzazione della cultura collaborativa alla ricerca di un'enciclopedia universale. Come progetto di collaborazione online, rappresenta un ottimo esempio di produzione sociale della conoscenza (rispecchiando gli interessi della popolazione degli utenti.) e lo studio degli agenti che producono il contenuto di Wikipedia è quindi un argomento di discussione di grande interesse sia per l'analisi delle reti che per le scienze sociali. A riguardo sono stati svolti diversi studi nel corso degli anni, seguendo approcci e metodologie differenti e sono stati ricavati anche risultati tra loro contrastanti. Wikipedia è caratterizzato da un'estrema dinamicità, si pensi solo che nel 2005 era definito tra i 100 siti più popolari [29], mentre ad Aprile 2014 è al sesto posto tra i siti

più visitati al mondo[6], per cui ci si rende conto come da un punto di vista temporale questa caratteristica implichi una grande variabilità dei risultati dei vari studi svolti nei diversi anni. I vari studi però hanno in comune il fatto di aver rilevato una struttura socio-produttiva che rispecchia l'esistenza di diverse categorie di utenti in base al numero di modifiche che questi applicano alle pagine. Si può rilevare sostanzialmente un gruppo minore di utenti *élite* che si differenzia da una maggioranza di utenti appartenenti alla categoria dei piccoli produttori. Queste due categorie sono associate alla produzione di contenuto differente anche in termini di *qualità*, nel senso che il gruppo *élite* si è mostrato associato alla produzione e modifica del 30% del reale contenuto, mentre gli utenti che effettuano poche modifiche tendono ad aggiungere perlopiù singole parole [27]. La distribuzione della produzione del contenuto da parte di questi macro-gruppi di utenti è stata mostrata essere una power law, in quanto è il gruppo *élite*, utenti che hanno eseguito più di 10.000 modifiche e che rappresentano meno del 5% del totale degli utenti, a produrre oltre il 50% del contenuto su Wikipedia, anche se nello studio di Kittur del 2007 viene mostrato come questa distribuzione vari nell'ultimo periodo di riferimento dello studio, passando dal 50% del 2002 al 20% del 2006 a causa dell'aumento di partecipazione da parte di piccoli produttori, intesi come utenti con meno di 100 modifiche, che arrivano a rappresentare quasi il 90% degli utenti attivi [27]. Il fondatore di Wikipedia, Jimmy Wales, nel 2005 affermò che è il 2% degli utenti a produrre il 75% del contenuto [38], anche se nello stesso periodo, lo studio di Swartz del 2006 [39] riporta una distribuzione più equa del lavoro; differenze dovute probabilmente alle diverse metodologie di analisi adottate.

In un lavoro del 2007, Priedhorsky introduce un modello di misurazione del *valore* delle modifiche inserite dagli utenti basato sul calcolo delle visualizzazioni delle pagine e delle singole parole, basandosi sul concetto che "non c'è valore in modifiche che non vengono visualizzate"[8]. Il valore di riferimento per definire il contenuto di una pagina di Wikipedia è il numero di volte che

[6]Statistiche Alexa, http://www.alexa.com/siteinfo/Wikipedia.org

viene visualizzata ogni parola introdotta da una modifica (PWV, Persistent Word Views): ogni volta che un articolo viene visualizzato, viene incrementato il valore del PWV di ogni parola dell'articolo. Di conseguenza gli autori che scrivono contenuto letto spesso forniscono un valore alla comunità e se un contributo è stato visto molte volte senza essere cambiato, è probabile che sia valido. Il punteggio finale di ogni autore è la somma di tutte le PWV di ogni articolo. Con questo studio Priedhorsky evidenzia come il dominio dell'*élite* degli autori persiste ed aumenta nel tempo, mostrando come lo 0,1% degli autori produca il 40% del *valore* di Wikipedia. Nello studio di Voss del 2005 si cerca di comprendere il motivo per cui, nel periodo oggetto dello studio, vi sia un calo nella percentuale delle modifiche inserite dagli *admin* di Wikipedia rispetto a quello dei primi anni successivi la nascita di Wikipedia. Questo comportamento è stato descritto non con un calo nell'attività degli admin, ma con un aumento del numero di modifiche da parte degli utenti *non-admin*, in particolare il declino della produzione degli utenti *élite* corrisponde ad un aumento di quella degli utenti con meno di 100 modifiche, la nuova *massa* di autori [29]. Questo andamento nella popolazione degli autori di Wikipedia, soprattutto localizzato nel periodo temporale preso in analisi da Voss, ovvero 2002-2006, rispecchia anche il tipico processo di adozione della tecnologia sul Web: gli utenti *élite* sono gli *early adopters* che hanno scelto per primi la tecnologia (Wikipedia) e la rifiniscono. In sistemi collaborativi online come Wikipedia però, vediamo come questi *early adopters* non si limitino a migliorare la tecnologia, ma vanno oltre, creando sufficiente valore nel sistema in modo che la massa di nuovi utenti potrà trovare un sistema a basso costo di partecipazione e al tempo stesso un modello di produzione del contenuto solido, basato su linee guida e procedure, verso cui il nuovo utente è attirato ad inserirsi e a partecipare, fornendo il proprio contenuto e facendo diffondere il sistema in maniera esponenziale. Queste misurazioni appaiono consistenti anche se applicate sulle versioni di Wikipedia in lingua non inglese, indicando quindi che deve esistere un processo generativo del contenuto, per cui una struttura socio-produttiva, coerente per tutti gli ambiti.

1.1.3 Perchè ci fidiamo di Wikipedia

In un articolo su Nature del Dicembre 2005, è stato riportato uno studio comparativo tra l'Encyclopaedia Britannica e Wikipedia, confrontate in termini di accuratezza delle pagine scientifiche [30]. Lo studio condotto dallo staff di Nature consisteva nel chiedere a accademici di analizzare 50 coppie di articoli estratti da Wikipedia e dal sito dell'Encyclopaedia Britannica, senza che questi sapessero da quale delle due enciclopedie proveniva ciascun articolo. Ciascun intervistato ha evidenziato una lista di errori per ciascun articolo, in totale 123 per la Britannica e 162 per Wikipedia. Ai commenti degli studiosi sugli errori sono stati quindi associati dei valori numerici dallo staff di Nature. In base a questi risultati, nello lo studio si conclude che le informazioni contenute su Wikipedia si possono considerare affidabili tanto quanto quelle della Britannica e che gli errori in Wikipedia sarebbero l'eccezione e non la regola. Questo risultato di per se non sminusice l'affidabilità della Britannica ma evidenzia l'alto livello di precisione di Wikipedia, caratteristica della quale non si sospettava fosse a livelli così elevati prima di questo confronto.

Questo studio però non è stato accettato dall'Encyclopaedia Britannica, che con un articolo di risposta ha contestato la validità dello studio di Nature, mostrando come fosse caratterizzato da bias e da una presentazione dei risultati fuorviante [31]. Viene messa in evidenza la discrepanza tra le conclusioni dello studio, in cui si afferma che Wikipedia si avvicina alla Britannica in termini di precisione dei suoi articoli scientifici, ed i dati contenuti all'interno dello studio stesso, leggendo i quali si nota che Wikipedia contiene un terzo di errori in più rispetto a quelli della Britannica, valore che secondo l'azienda inglese non può portare ad una simile conclusione. Molte delle segnalazioni effettuate dagli intervistati riguardo gli articoli della Britannica non sono stati accettati dalla Britannica stessa. Sull'argomento è nata quindi una diatriba, seguita dalla controrisposta di Nature [37] che ha sostenuto e confermato i risultati dello studio originale. A distanza di sette anni dallo studio condotto da Nature, Wikimedia Foundation ha commissionato uno studio al-

l'azienda inglese di e-learngin Epic e all'Università di Oxford per un nuovo confronto[35], che tenesse conto della forte evoluzione che l'enciclopedia online ha subito negli anni, passando dai quasi 4 milioni di pagine al tempo dell'articolo di Nature, agli oltre 23 milioni del 2012. Differenza del nuovo studio rispetto a quello precedente è stato il fatto che le 22 pagine scientifiche sottoposte all'esame degli accademici di madre lingua appartenessero non solo a Wikipedia in lingua inglese, ma anche spagnolo e arabo. L'accuratezza delle pagine di Wikipedia nelle 3 lingue è stata quindi confrontata con Encyclopaedia Britannica, Enciclonet e Arab Encyclopaedia. Secondo le analisi Wikipedia vincerebbe su tutti i fronti: stile, numero di fonti, all'accuratezza delle notizie e ovviamente velocità di aggiornamento[34].

Tralasciando i dettagli dal dibattito sopracitato, i risultati degli studi di Nature e dell'Università di Oxford rappresentano un importante elemento per poter introdurre l'argomento attendibilità di Wikipedia, che ci porta a discutere della fiducia degli utenti verso l'enciclopedia libera. Wikipedia stessa afferma che basare l'esistenza dell'*enciclopedia libera* sulla sola attendibilità sarebbe molto complesso: infatti Wikipedia non può assicurare l'attendibilità dei propri testi e utilizza quindi un modo diverso di concepire e condividere il sapere rispetto alle enciclopedie tradizionali. Non sono quindi i testi in sé ad essere direttamente attendibili, come nelle enciclopedie tradizionali, ma piuttosto le fonti attendibili e le fonti verificabili utilizzate per scriverli: approccio che di norma rende attendibili i testi stessi. Queste caratteristiche portano Wikipedia ad essere "non un'altra enciclopedia, ma una nuova enciclopedia" [46]: un'enciclopedia tradizionale può fornire una notizia attendibile, ma che comunque appare cristallizzata nelle pagine, che non ammette discussione e non ammette confronto, in quanto fornisce un sapere che viene dall'alto. Wikipedia non può essere un'enciclopedia sempre attendibile in senso tradizionale: vuole infatti riformulare il concetto stesso di attendibilità, sostituendolo piuttosto con quello di verificabilità. In effetti, a Wikipedia non interessa stabilire quale sia la Verità assoluta; il suo scopo è piuttosto dare un quadro completo delle diverse opinioni perché il lettore

possa decidere con la propria testa, e naturalmente integrare la voce sulla base di altre fonti. Per questo uno dei pilastri su cui si basa Wikipedia è la presenza del punto di vista neutrale.

Come bisognerebbe quindi usare Wikipedia? Wikipedia non viene usata come una enciclopedia tradizionale, ma sapere di non doverla pendere come un'autorità definitiva non è sufficiente per dire quanto e come dovremmo effettivamente basare le nostre convinzioni su di essa.

P.D. Magnus, professore associato del Dipartimento di Filosofia presso l'Università di Albany, NY, in un articolo del 2009 descrive i 5 criteri tramite i quali le persone valutano le affermazioni nella vita reale, e per ciascuno di questi ne mostra il funzionamento rispetto a Wikipedia [10]:

- Autorevolezza
- Plausibilità dello stile
- Plausibilità del contenuto
- Calibrazione
- Campionatura

Autorevolezza Si crede in una affermazione perchè questa proviene da una fonte attendibile, il che sposta direttamente il problema sul decidere quali fonti lo sono e quali no. Ma per decidere se una fonte è attendibile ci si basa sulle conoscenze di background che si possiedono riguardo la fonte stessa, ad esempio del sito del New York Times ci si fida come del giornale, di un utente di un forum ci si fida come lo si farebbe di una persona, se la si conosce personalmente o si conosce chi si fida di quest'ultima. Per Wikipedia in particolare, conoscere chi scrive non è importante, come non lo è conoscere l'autore di un articolo sul New York Times. La sua autorevolezza deriva da Wikipedia. Wikipedia invita a usare le fonti, ma le fonti non sono verificabili in senso pratico poichè verificare una fonte, specialmente se non online, annullerebbe il vantaggio ottenuto dall'utilizzo Wikipedia. Ma l'affidabilità

di Wikipedia varia molto da articolo ad articolo, per cui dire che Wikipedia sia autorevole è una posizione troppo generica. Studi hanno proposto di valutare l'affidabilità in base agli argomenti (ad esempio scienze o filosofia) ma non è sufficiente per distinguere i singoli articoli, per i quali l'affidabilità è variabile. In blog di esperti vengono riportati i link a Wikipedia, ma il fatto che il link sia ritenuto corretto dall'esperto non vuol dire che quando l'utente raggiunge quel link, la pagina sia uguale a come quando è stata letta dall'esperto che l'ha riportata ritenendola valida.

Plausibilità dello stile In generale se qualcuno che afferma di avere un PhD scrive con lo stile di un ragazzo del liceo, nel lettore si crea un giudizio prima ancora del valutare la plausibilità delle affermazioni che vengono espresse con quello stile. Ma possedere capacità espositive non equivale ad essere una fonte attendibile, anche se è spesso associata a conoscenza dell'argomento trattato. Nello specifico di Wikipedia, un utente può essere capace di correggere errori grammaticali e di struttura della frase inserendo il wiki-markup corretto senza possedere alcuna conoscenza dell'argomento dell'articolo che sta correggendo. Inoltre, anche se l'utente conoscesse qualcosa dell'argomento, questo non garantisce che si impegnerà a determinare se ci sono affermazioni false all'interno dell'articolo.

Plausibilità del contenuto In generale si tende a non credere in una affermazione, anche se chi la riporta è una fonte attendibile, quando il contenuto è palesemente errato. In Wikipedia la ricerca della plausibilità ci porta fuori strada maggiormente rispetto al caso di una singola fonte autorevole semplicemente perchè altre persone hanno consultato precedentemente la pagina ed hanno rimosso le affermazioni palesemente errate, ma questo non implica che i contenuti più specifici siano corretti ed affidabili.

Calibrazione Un'affermazione non accettata presa così com'è, ma viene confrontata con le conoscenze che già si possiedono a riguardo. Se le nuove informazioni sono in linea rispetto alle conoscenze che l'utente già possiede,

allora l'affermazione verrà ritenuta attendibile per ragionamento induttivo (è giusta sui punti che conosco, allora sarà giusta anche sugli altri). É come l'*autorevolezza*, ma senza possedere conoscenze antecedenti sulla fonte, bensì sull'argomento. La calibrazione però non funziona se quello che l'utente può verificare non è rappresentativo delle affermazioni fatte dalla fonte. Quando si applica la calibrazione su un articolo di Wikipedia, è necessario che un utente conosca già qualcosa riguardo l'argomento trattato oppure abbia una fonte alternativa da consultare. Tuttavia, perchè la calibrazione sia efficace, il range di affermazioni che un utente può controllare non deve essere semplicemente il range di affermazioni che molti altri wikipediani possono controllare. Anche se questo è vero per quanto riguarda la popolazione di utenti che contribuiscono alla produzione di contenuto, difficilmente si può sapere se questo sia vero in assoluto. Di conseguenza la calibrazione fallisce quando si applica agli articoli di Wikipedia.

Campionatura (*Sampling*) La campionatura si può vedere come la richiesta di un'altra opinione e, se nel caso in cui questa coincida, si porrà fiducia sull'affermazione. Se si consulta unicamente un articolo di Wikipedia, allora non ci si preoccupa di campionare. Supponiamo però che qualcuno allarghi la sua ricerca guardando anche da qualche altra parte sul web; molte informazioni contenute nelle pagine sono raccolte da Wikipedia e riportate senza specifiche evidenze temporali della fonte. Wikipedia è dinamica, per cui una pagina web creata un anno fa citando Wikipedia può contenere riferimenti ad affermazioni presenti in quel momento senza che però la pagina sia una copia effettiva dell'articolo. Quindi, anche escludendo le copie identiche, gli utenti di Wikipedia possono trasmettere affermazioni da fonti conosciute verso Wikipedia e da Wikipedia verso le loro pagine Web. Quindi con il campionamento può capitare di considerare "altre opinioni" affermazioni che erano presenti in un articolo di Wikipedia.

Come si è visto, applicare su Wikipedia i criteri naturali di valutazione dell'affidabilità delle affermazioni risulta spesso inadeguato. Non sono perciò

questi i parametri utilizzati dagli utenti che consultano Wikipedia per potersi fidare del contenuto dell'enciclopedia online.

Perchè ci si fida quindi di Wikipedia? In letteratura sono presenti studi che esaminano le ragioni per cui gli utenti si fidano di Wikipedia. Secondo lo studio di Jean Goodwin [3], consultando Wikipedia l'utente pone fiducia sulle affermazioni scritte per le stesse ragioni che lo portano a fidarsi dell'opinione degli esperti in generale:

1. sulla base della competenza dei suoi autori individuali
2. sulla base della conoscenza collettiva che emerge dalle interazioni di molti autori
3. sulla base dell'esperienza passata sulla sua affidabilità

Goodwin mostra come anche le suddette ipotesi in realtà non siano applicabili su Wikipedia.

Per quanto riguarda la fiducia nella competenza dei singoli, relativamente a Wikipedia questa analisi suggerisce che gli utenti consultino il sito in base alla conoscenza di coloro che vi contribuiscono, contando sul fatto che gli editor sappiano ciò di cui stanno parlando. Appare tuttavia ovvio come tale approccio non possa funzionare per almeno tre ragioni:

- nonostante la registrazione di tutte le modifiche fatte a ogni articolo, di norma è impossibile risalire da tali modifiche a una persona specifica, di cui si possa valutare la conoscenza al di fuori dei suoi contributi sul sito;
- anche coloro che si registrano non sono tenuti a farlo col loro vero nome, e l'uso degli pseudonimi è assai frequente;
- Wikipedia ha deciso di non utilizzare tecniche di verifica delle credenziali inserite dagli utenti registrati

Riassumendo, l'anonimato o la pseudonimia degli editori di Wikipedia impediscono di valutarne la competenza; quel poco che se ne può ricavare non

fornisce particolare confidenza sulle loro conoscenze e la cultura anti-esperti di Wikipedia non dà ragione di credere che tale stato di cose sia destinato a cambiare [50]. Dunque, le competenze degli autori individuali non sembrano in grado di giustificare gli utenti a consultare Wikipedia.

Per quanto riguarda la saggezza delle folle, secondo il ragionamento di Goodwin appare come le qualità epistemiche degli autori, individuali o collettivi che siano, non possano giustificare la prassi di consultare Wikipedia. Le ragioni principali che portano a questa conclusione derivano dalle analisi sulla struttura produttiva dalla quale sembrerebbe che Wikipedia esprima la saggezza (ammesso che di saggezza si tratti) di un'élite, piuttosto che di una folla (power law della produzione degli articoli). Inoltre facendo riferimento al momento esatto della consultazione, una voce di Wikipedia non è il prodotto dell'intelligenza di un *alveare* o di uno *sciame*, ma piuttosto dell'ultima intelligenza che vi ha posto mano. Teoricamente gli articoli di Wikipedia sono il risultato di un lungo processo argomentativo che avviene sulla pagina di *Discussione* associata ad ogni voce, ma ciò che appare in un articolo di Wikipedia non sempre è il risultato di un processo di negoziazione; invece, può essere semplicemente ciò che il curatore più recente ha deciso di inserivi o di lasciarvi. L'ultima persona ad apportare modifiche al testo prima che venga consultato potrebbe aver deciso di cestinare l'intero articolo.

Per quanto riguarda l'esperienza passata, ovvero il fatto che si continua a consultare Wikipedia perché è stata trovata utile e affidabile in passato, Goodwin afferma che è improbabile che l'utente si fermi ad esplorare se i dettagli offerti da Wikipedia siano corretti quando consulta una pagina, in quanto sta cercando informazioni su un argomento di cui conosce poco, per cui il suo affidamento a Wikipedia non si può definire giustificato dall'esperienza

Nessuno di questi criteri epistemici fornisce una spiegazione sul perché gli utenti si fidino di Wikipedia, invitando piuttosto piuttosto ad essere agnostici [3]. Un paradosso insito in ogni tentativo di valutare l'autorevolezza degli esperti su basi epistemiche è che si consulta un esperto perché si è ignoranti sul dominio di cui l'altro è esperto. Ma per valutarne la competenza,

è necessario giudicarne le conoscenze, e ciò implica non essere ignoranti sul dominio in questione. Goodwin propone quindi di rispondere alla domanda sulla fiducia verso Wikipedia tramite un approccio di valutazione pragmatico delle azioni dell'esperto, e non più sulla valutazione epistemica della sua competenza. Ci si chiede se quell'artigiano abbia una buona reputazione da difendere nella nostra comunità, o quali garanzie ci offra della bontà del suo lavoro. Se ci fidiamo di lui, ce ne fidiamo in termini pragmatici, ovvero ci fidiamo del fatto che abbia le abilità che promette di usare al nostro servizio. Le domande critiche per verificare un appello all'autorità di un esperto potrebbero includere le seguenti: *"Perché mi sta offrendo la sua opinione?", "Posso verificare le sue intenzioni?", "Cosa ha da perdere, se scoprissi che si sbaglia?", "Ci sono modi di garantire che paghi tali penalità, in caso di errore?"*. Tutti possiedono quindi dei metodi per valutare ciò che un uomo della strada, un esperto, o Wikipedia afferma, anche quando non si possiede alcuna competenza sull'argomento in questione.

In definitiva quindi, perché è ragionevole per l'utente consultare Wikipedia? Secondo le conclusioni dell'articolo di Goodwin il motivo è che i Wikipediani si preoccupano profondamente della qualità del loro lavoro[3]. Ma per fidarsi di loro sulla base della passione dedicata al progetto, i Wikipediani dovrebbero anche offrire ulteriori rassicurazioni: in particolare evidenze del fatto che tale entusiasmo non produca predicibili distorsioni, errori ed omissioni. Come può allora l'utente essere sicuro che ciò che trova su Wikipedia sia stato scritto da una persona motivata dall'ideale di cui sopra? Infatti, non sembra esserci nulla che impedisca a qualcuno di modificare una voce per interesse personale (ad esempio, per promuovere le proprie idee o prodotti) o per dispetto (vandalismo), tant'è che nel pieno dell'approccio *open*, non viene richiesta neppure una registrazione nel sistema per poter modificare i contenuti. I lettori quindi devono fare affidamento alle linee guida sviluppate da Wikipedia per la produzione del contenuto e per la protezione delle pagine, ovvero fidarsi che chi scrive sia"in buona fede" e che segua precisamente le linee guida, supponendo che queste siano effettivamente sufficienti

per garantire la qualità delle informazioni e che i sistemi di gestione dei vandalismi siano sufficienti a mantenere valide le informazioni contenute nelle pagine. Nella sezione 1.3 di questo capitolo saranno descritti in dettaglio i metodi utilizzati da Wikipedia per proteggere le informazioni contenute nelle pagine, mentre nei successivi capitoli saranno riportati i risultati degli esperimenti sviluppati in questo studio per comprendere se questi presupposti per la fiducia possano considerarsi sufficienti.

1.2 La struttura di Wikipedia vista come un network

1.2.1 Wikipedia in termini di rete

Dall'analisi delle reti sociali emergono di frequente modelli core-periphery che descrivono diversi aspetti delle dinamiche d'interazione tra gli attori, come la diffusione della conoscenza, la distribuzione del lavoro (sia a livello aziendale che inter-aziendale), la struttura lavorativa aziendale, l'adozione delle nuove tecnologie. Un esempio è lo studio delle relazioni tra membri appartenenti a differenti gruppi di lavoro basato sulla mappatura dei flussi di informazione attivi fra questi. La rilevanza dei membri nella struttura aziendale viene rispecchiata da una topologia di rete di tipo core-periphery, in cui la periferia comprende i soggetti meno attivi nel gruppo e il nucleo invece include i membri chiave [9]. Anche lo studio delle perfomance all'interno dei gruppi di lavoro è stato dimostrato rappresentabile con modelli core-periphery [19]. Questa topologia di rete è stata usata anche per descrivere alcune strutture lavorative aziendali, con un core caratterizzato da impiegati stabili, ad alta esperienza e con percorsi di carriera interni definiti e una periferia di soggetti meno inseriti nella struttura aziendale e diversificati rispetto ai primi per vari aspetti (importanza delle attività svolte, ruoli ricoperti o caratteristiche lavorative non coincidenti con quelle globali dell'azienda come l'orario di lavoro o il tipo di inquadramento) [18]. Queste

analisi rendono le interazioni "visibili" e quindi applicabili in pratica nell'organizzazione aziendale. Per quanto riguarda la diffusione della conoscenza, è stato mostrata l'esistenza di uno schema core-periphery nella rete di diffusione delle tecniche di lavoro tra agricoltori di alcuni villaggi africani: gli attori appartenenti al core, oltre che portare una maggior diffusione dell'informazione, sono più attivi nella raccolta di nuove informazioni dall'esterno rispetto agli attori periferici [17].

Anche la struttura di Wikipedia può essere descritta in termini di rete, sia da un punto di vista strutturale, per quanto riguarda i collegamenti tra le pagine, che da un punto di vista socio-produttivo, ovvero degli utenti che contribuiscono alla creazione del contenuto.

In termini strutturali, un modello tramite il quale rappresentare la struttura di rete di Wikipedia è il modello core-periphery, in cui il nucleo, costituito da pagine raggiungibili da un altissimo numero di percorsi differenti, è circondato da una periferia formata da molte pagine con pochi collegamenti tra loro ma che si collegano a qualche elemento del core, attraverso il quale diventa possibile raggiungere ogni pagina della rete. Nei rispettivi lavori effettuati nel 2006, Capocci e Buriol hanno evidenziato come il grafo di Wikipedia rappresentante i collegamenti (*hyperlink*) tra le pagine, che costituiscono i nodi del grafo, sia identificabile nello specifico con un modello *bow-tie*, associabile a quello del Web, anche se con un processo di sviluppo differente da quest'ultimo [14, 13]. In questi studi si mostra come in termini strutturali Wikipedia possegga un denso *core*, composto da pagine molto ricche di link in entrata e in uscita; l'82,7% delle pagine in Wikipedia in versione italiana è contenuto nel *core* [14].

Oltre che dal punto di vista strutturale, Wikipedia è rappresentabile in termini di rete anche per quanto riguarda lo schema socio-produttivo sottostante, in cui i nodi del grafo sono rappresentati dagli utenti e i collegamenti tra questi sono le pagine per cui hanno inserito una modifica. Anche in questi termini, il modello che rappresenta questa rete è ipotizzabile essere un modello core-periphery: è presente un *core* composto da un numero limitato di

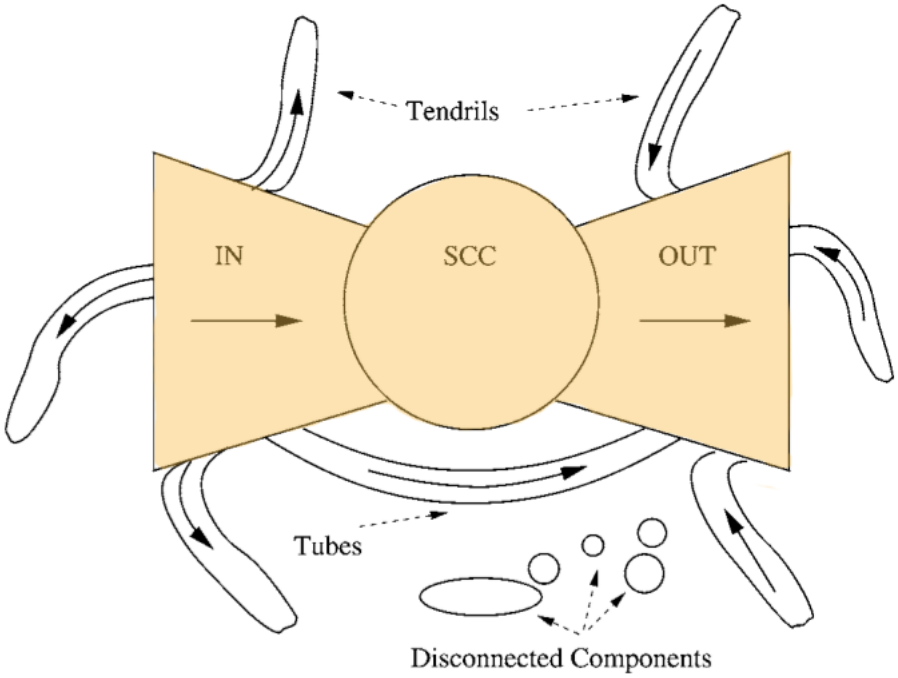


Figura 1.5: Modello Bow-tie come rappresentato da Capocci [14]. La percentuale di pagine nelle aree Tendrils e Tubes si possono definire trascurabili

utenti che eseguono contributi su molte pagine e di conseguenza possiedono un alto gradi di correlazione tra loro, che sono connessi a loro volta con una *periphery* composta da un alto numero di utenti che eseguono un numero limitato modifiche e sono poco o per niente correlati tra loro. Ancora in termini di schema socio-produttivo, questa topologia di rete si ipotizza rilevabile anche se si considerano come nodi le pagine e i collegamenti tra questi sono costituiti dagli utenti ad esse comuni: il nucleo sarà costituito da quelle poche pagine che possiedono molti utenti in comune, la periferia dalle molte pagine connesse ad altre tramite pochi utenti comuni.

1.2.2 Proprietà di rete

Le power law

In ambiti come la biologia e le scienze naturali, le variabili sono generalmente rappresentate da distribuzioni Gaussiane. Per quanto riguarda invece le relazioni tra le entità in gioco in ambiti come la fisica, le scienze sociali, l'informatica, l'economia, queste seguono spesso distribuzioni *power law*, adatte a modellizzare ad esempio la popolazione delle città, per le quali si nota come la dimensione di una città sia inversamente proporzionale alla sua posizione nell'ordinamento decrescente delle città per dimensione , ovvero che la dimensione della città che occupa la posizione 100 della classifica è 1/10 della dimensione della città che occupa la posizione 10 della stessa classifica. Questo è un tipico esempio di distribuzione di Zipf [2], che fu usata inizialmente come modello per la misurazione della frequenza delle parole all'interno di testi scritti, ma ora è associata anche agli oggetti del mondo web. La distribuzione di Zipf è una distribuzione discreta la quale definisce che quando gli elementi sono catalogati in ordine di popolarità discendente, la frequenza degli elementi è inversamente proporzionale alla loro posizione nell'ordinamento decrescente rispetto alla frequenza stessa. Questa distribuzione può essere considerata una rappresentazione discreta della distribuzione di Pareto, ottenuta trasformando gli assi della distribuzione di Pareto. Altri esempi di variabili rappresentabili attraverso power law sono la distribuzione della ricchezza nella società, che segue il principio di Pareto, la "regola del 80-20" per cui il 20% della popolazione possiede l'80% della ricchezza (esempio classico riportato in 1.6), così come la distribuzione delle citazioni degli articoli scientifici, il numero di accessi ad un sito web o le performance di una rete LAN [16].

Una manifestazione delle distribuzioni power law sono le cosiddette "long tail", che esemplificano la proprietà statistica per cui esistono molti più eventi poco frequenti rispetto ad una distribuzione Normale. La long tail è stata associata ad esempio alla frequenza delle parole chiave utilizzate per le ri-

x	N	x	N	x	N
150	400,648	600	42,072	2,000	9,880
200	234,185	700	34,269	3,000	6,069
300	121,996	800	29,311	4,000	4,161
400	74,041	900	25,033	5,000	3,081
500	54,419	1,000	22,899	10,000	1,104

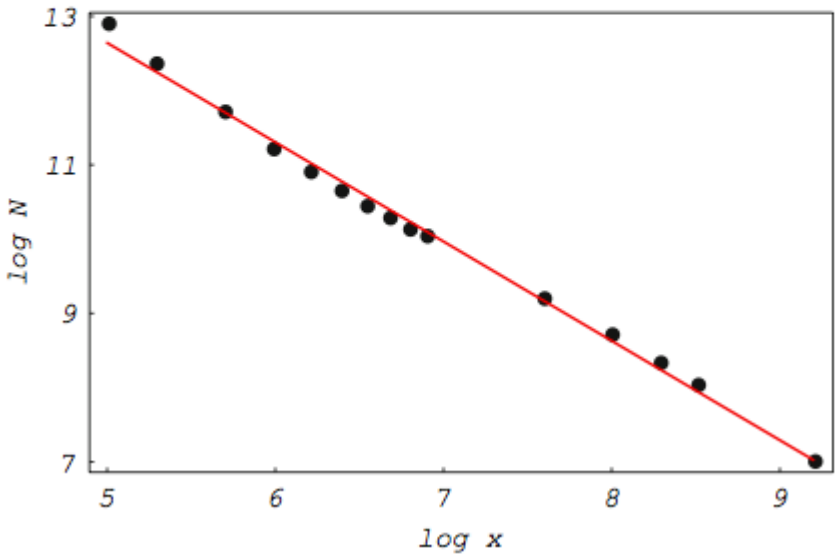


Figura 1.6: Esempio legge di Pareto riportata in [2]. Rappresenta la distribuzione del reddito di una categoria di contribuenti inglesi per gli anni 1893-1894. x rappresenta il reddito annuale e *N(x)* il numero di contribuenti il cui reddito annuale è maggiore di x

cerche; questo significa che se si ordinano le parole chiave dalla più diffusa alla meno diffusa, in base a quante volte una parola chiave è stata utilizzata, troviamo che ci sono solo poche parole chiave che vengono usate spesso mentre c'è una lunghissima "coda" di parole usate raramente. Poiché la lista di queste parole chiave è molto grande, le parole non frequenti rappresentano una percentuale rilevante all'interno del totale [15].

In uno studio del 1999, Faloutsos[21] ha identificato che la struttura di Internet è caratterizzata da una power law, nel senso che la distribuzione di

probabilità della connettività di un nodo segue questa legge. É stato osservato che nelle comunità online, generalmente è una minoranza del gruppo dei partecipanti quella che produce la maggior parte del contenuto, mentre la maggior parte degli utenti produce poco (o nessun) contenuto, seguendo quindi una legge power law. Queste dinamiche sono state osservate anche per i blog [33]. Vi sono diversi esempi dalla letteratura che mostrano come anche la creazione e fruizione del contenuto su Internet segue la legge power law, con pochi utenti che creano contenuto e moltissimi che ne fruiscono senza collaborare attivamente. Uno studio del 2007 di Gill mostra come la popolarità dei video di YouTube segue la distribuzione di Zipf, anche se con differenze nell'andamento della coda [23]. Le caratteristiche tipiche della power law sono state identificate anche in altri servizi di condivisione di video prodotti dagli utenti. La stessa distribuzione è stata rilevata anche in altre strutture legate ad Internet, come ad esempio il numero di *amici* per ciascun utente nei social network [24] o la popolarità dei canali sulle Internet TV, con l'80% degli spettatori collegato sul 10% dei canali.

Wikipedia, power law e invarianza di scala

Wikipedia è uno dei prodotti più importanti del Web[40], per cui ci si aspetta che questo segua le stesse distribuzioni che la letteratura ha evidenziato per molti degli altri prodotti di Internet, ovvero le power law, e così è sicuramente stato nei primi periodi dopo la nascita di Wikipedia [25, 26, 28].

Per quanto riguardo la rete che rappresenta la struttura di Wikipedia, è stato mostrato come la distribuzione dei link in entrata verso il *core* possegga una distribuzione power law, mentre per i link in uscita si nota una distribuzione log-normale[13]. Nello studio di Buriol viene mostrato come il valore ricavato per l'esponente $\gamma = 2.00$ sia associabile ai valori di collegamenti in entrata del *Webgraph*, rete descrittiva della struttura generale del Web, in cui i nodi sono rappresentati da pagine statiche ed i collegamenti tra questi sono rappresentati dagli hyperlink tra queste [36].

Per quanto riguarda la produzione del contenuto e la struttura socio-produttiva, come si è detto precedentemente, sono presenti in letteratura diversi studi che mostrano come la produzione del contenuto avvenga seguendo una distribuzione power law, in cui è una minoranza di utenti *élite* a produrre la maggior parte del contenuto e la restante maggioranza contribuisce invece effettuando poche modifiche [8, 29, 27, 38]. La power law è però rilevabile anche per quanto riguarda la distribuzione delle modifiche sulle pagine, per cui il 53% delle pagine ha ricevuto più di 10 modifiche, e circa il 5% più di 100 modifiche [13]. Buriol spiega questo comportamento innanzitutto con il fatto che la frequenza di modifica segue la distribuzione dell'importanza della pagina, per cui una pagina più è visualizzata, più sarà oggetto di modifche. Nello stesso studio si evidenzia come anche la distribuzione del numero di utenti distinti che contribuiscono ad un articolo segua la power law, per cui il 50% degli articoli sono stati modificati da più di 7 differenti autori, mentre solo il 5% da oltre 50 autori. Anche Voss con uno studio basato sul calcolo del numero di modifiche [29], mostra la crescita esponenziale avuta dal contenuto di Wikipedia dal 2002 e una distribuzione del numero degli autori unici che segue una power law, così come quella del numero di articoli per autore. In quanto progetto collaborativo, tutti sono invitati a modificare gli articoli, i quali di conseguenza possono arrivare ad avere anche un grande numero di autori. Anche all'interno dello stesso articolo quindi, la distribuzione del numero di autori distinti per ogni articolo è una power law. Voss mostra come per articoli con un numero tra 5 e 40 di autori distinti, il numero di autori segue una power law con esponente 2.7 [29].

Come dimostrato per reti complesse come il Web, i social network o le reti relative alle e-mail inviate tra gli utenti [20], anche Wikipedia possiede la caratteristica di essere *scale-free*, in quanto è possibile rilevare la presenza dello stesso comportamento indipendentemente dal frammento di prodotto che si prende in analisi[14]. Nel caso specifico è sempre rilevabile una distribuzione power law per le variabili di interesse, indipendentemente dalla

scala di riferimento. Come sarà descritto in dettaglio nel prossimo capitolo, è possibile notare una distribuzione power law riferita al numero di modifiche per utente, sia a livello di singola categoria di pagine, sia a livello generale dell'intero sistema.

Queste caratteristiche portano a ipotizzare l'esistenza di sistemi di autocorrezione dipendenti dalla tipologia di pagina, ovvero dalla sua posizione nella rete, in quanto si può ipotizzare che la probabilità per una pagina di essere corretta sia inverso al valore di scale-free relativo: più una pagina è correlata alle altre o quanto più questa è comune tra gli utenti che applicano le modifiche (appartenendo cioè ai *core* dei grafi), tanto maggiore ci attende sia la probabilità che un errore inserito in quella pagina venga rilevato e corretto.

1.3 Wikipedia in action: aggressioni vandaliche e capacità di resilienza

1.3.1 Gestione degli errori in Wikipedia

Come già detto, la caratteristica chiave di Wikipedia, la completa libertà di modifica da parte di chiunque, comporta alcuni rischi per l'utente che consulta gli articoli, che può trovarsi a leggere informazioni non verificate, inesatte o addirittura volutamente erronee, che possono rimanere sulle pagine di Wikipedia per diverso tempo. Per questo motivo Wikipedia non dà garanzie sulla validità dei contenuti presenti nelle pagine del sito, per i quali all'interno del sito stesso si afferma che non è possibile garantire una verifica o un controllo da parte di soggetti legalmente abilitati o con le necessarie competenze per esprimersi nei campi trattati. E' Wikipedia stessa a definire che un sarebbe necessario proprio un controllo di questo tipo per fornire un'informazione completa, corretta e certa [45] e a mettere in guardia gli utenti con apposite pagine di avvertenze e disclaimer[1]. Alcune edizioni di

[1]Relative pagine: Disclaimer sul rischio, Disclaimer legale, Disclaimer medico

Wikipedia hanno adottato sistemi per selezionare e approvare le versioni più accreditate, ma sempre in forma volontaria e senza garanzia. La cosa più vicina a questo sistema di selezione è l'attuale sezione *vetrina*, ma anche le voci citate in essa potrebbero essere state modificate senza il controllo necessario. Sebbene vi siano delle regole e un sistema non ufficiale di gestione anonimo e volontario, Wikipedia non è uniformemente revisionata; anche se gli utenti possono correggere errori, non hanno tuttavia alcun dovere di farlo e quindi tutte le informazioni presenti sono fornite senza alcuna garanzia di idoneità per qualsiasi scopo o utilizzo.

Wikipedia agisce in maniera differente per preservare la qualità delle informazioni già esistenti che vengono modificate, in base che si tratti di informazioni errate inserite in modo maldestro o impreciso durante un tentativo di miglioramento di una voce o che si tratti piuttosto di informazioni errate inserite volutamente per danneggiare la pagina, cosiddetti vandalismi. La gestione degli errori appartenenti alla prima categoria è necessariamente spostata sugli utenti: da una parte, in maniera preventiva, su quelli che contribuiscono a produrre il contenuto di Wikipedia, e dall'altra, in una sorta di scarico di reponsabilità, sugli utenti che consultano le pagine. Sono state sviluppate delle linee guida che ricordano agli utenti che modificano i contenuti di Wikipedia dell'importanza di indicare esplicitamente le fonti dalle quali sono tratte le informazioni riportate, per permettere così a chiunque legga gli articoli di individuare in maniera autonoma l'origine delle informazioni e di verificarne validità e attendibilità, ovvero accertarsi che quanto afferma il testo sia già stato pubblicato da una fonte attendibile. Wikipedia sottolinea a riguardo il termine *verificabilità*, distinguendolo da *verità*: un'informazione verificabile può anche essere falsa, un'informazione non verificabile può anche essere vera. Tuttavia in genere la verificabilità è un buon criterio di verosimiglianza di un'informazione [47]. Wikipedia quindi da un lato non vincola in nessun modo l'inserimento di nuove informazioni, mantenendo il massimo grado di libertà nella produzione del contenuto e invitando i Wikipediani ad essere diligenti e citare le fonti, dall'altro ricorda che non ha possibilità di

garantire la veridicità delle informazioni per cui sposta l'onere della verificabilità sui lettori. Inoltre anche la non completa conoscenza dell'argomento da parte dei contributori può causare l'inserimento di informazioni errate nelle pagine. In genere, errori di questo tipo vengono risolti a medio termine, perché prima o poi qualcuno che conosce l'argomento raggiunge la pagina e la corregge. Nel breve termine, alcuni ritengono che Wikipedia sia più autorevole dei quotidiani (che spesso contengono errori marchiani) ma meno di un'enciclopedia, almeno per le voci con una cronologia scarsa [47, 30].

Per quanto riguarda invece la gestione dei vandalismi, Wikipedia ha sviluppato sistemi di controllo e supervisione che vanno oltre suggerimenti, politiche, linee guida e pareri, il cui scopo è definire per la comunità ciò che sia contenutisticamente appropriato in pagine particolarmente esposte a rischi di vandalismo (es. biografie), oltre che distinguere il confine fra autopromozione e informazione legittima, ma puntano piuttosto su azioni di contrasto dei vandali. Si definisce vandalismo l'aggiunta, la cancellazione o la modifica, spesso reiterata, di contenuti e dati fatta con un evidente interesse o una malafede e con il conseguente risultato di compromettere l'integrità di Wikipedia. Di seguito un elenco di alcune forme di vandalismo registrate da Wikipedia [49]:

- Inserimento di informazioni volutamente errate
- Spamming (l'aggiunta di collegamenti esterni a fini promozionali o a siti non consentiti)
- Aggiunta di parole senza senso o cancellazione non giustificata di parti del testo
- Inserimento di insulti o ingiurie

Wikipedia invita tutti gli utenti, qualora individuino un vandalismo, a rimuoverlo immediatamente, eventualmente contattando un amministratore per i casi in cui occorrano azioni non consentite all'utente (come il blocco della pagina). Così come è possibile per chiunque inserire un vandalismo, allo

stesso modo è semplice rimuoverlo: è sufficiente visualizzare la cronologia della pagina, cliccare il campo in cui compaiono data e ora della modifica della versione precedente a quella vandalizzata, andare in modifica, indicare nell'oggetto che si tratta di un rollback per vandalismo e quindi salvare la pagina. In alcuni casi è possibile annullare rapidamente la versione vandalizzata agendo sul campo apposito, (annulla). Nei casi di insulti, blasfemie o informazioni la cui presenza contrasta con la politica di wikipedia in materia di privacy, l'utente è invitato a contattare qualche amministratore per la rimozione del vandalismo dalla cronologia, nella pagina delle richieste agli amministratori o utilizzando il template RichiestaPulizia. Nonostante i vandalismi siano di forte impatto negativo su Wikipedia, questa invita gli utenti [48] ad una gestione dei vandalismi attiva ma al tempo stesso non aggressiva, con un approccio che permetta agli utenti vandali di correggere il loro comportamento, ovvero prima di inserire un utente nella lista degli utenti da bloccare, è importante che: *"[...]sia ben chiaro cosa si intende per vandalismo, il vandalo sia stato avvertito, e nonostante ciò perseveri nello stesso genere di condotta, i vandalismi siano recenti e che il vandalo non abbia interrotto il proprio comportamento dopo essere stato avvisato".* Queste regole di condotta sostanzialmente chiedono ai Wikipediani di dare per scontato che ogni nuovo arrivato nella comunità sia in buona fede e dunque accoglierlo e fare il possibile per convertirlo quanto prima alla cultura interna al progetto centrale nelle attuali linee guida, laddove si chiede agli utenti di "non mordere i nuovi arrivati, per cortesia"[2]. Per dare coerenza a questa idea, un gruppo auto-organizzato di Wikipediani si è preso la responsabilità di agire come comitato di benvenuto. Benché questo approccio basato sul *WikiLove* possa evitare vandalismi dovuti a incuria o ignoranza, di nuovo servirà a poco nel combattere utenti motivati da reale malevolenza o interesse personale. Oltre a questo approccio di "autogestione" dei vandalismi da parte dei comuni utenti di Wikipedia, vi è un'ulteriore linea di difesa: gruppi auto-organizzati (chiamati *pattuglie*) che si assumono la responsabilità di individuare e cor-

[2]Wikipedia: Please do not bite the newcomers

reggere interventi impropri sul sito - Recent Changes Patrol o New Pages Patrol. I loro compito è monitorare le nuove modifiche a voci pre-esistenti e la creazione di nuove pagine, badando a individuare e correggere ogni vandalismo. Tale pattugliamento veniva originariamente condotto manualmente, ma la rapida crescita di Wikipedia ha spinto i Wikipediani a sviluppare una serie di strumenti software per automatizzare parte del processo (Wikipedia: Recent changes patrol), consentendo così di eliminare rapidamente la maggior parte delle modifiche malevole[8]. Nel 2009 è stato introdotto uno strumento ancor più sofisticato, in grado non solo di identificare ma anche di correggere in modo semi-automatico i casi più palesi di abuso [3]. Ma nonostante queste politiche e strumenti di controllo, come si afferma nella pagina relativa all'Attendibilità di Wikipedia, il vandalismo è il punto più debole del modello di Wikipedia [46]. Mentre le parolacce aggiunte alle pagine sono poco importanti (perché vengono facilmente notate e cancellate), esistono vandalismi più sottili, che riguardano informazioni molto specifiche e che per questo potrebbero passare inosservati per molto tempo. Al momento non vi sono soluzioni definitive al problema.

Il caso del *Brazilian aardvark* è un esempio emblematico del rischio in cui può incorrere il lettore di Wikipedia. Nel 2008, un 17enne newyorkese al ritorno da un viaggio in Brasile, ha inserito su Wikipedia un'informazione riguardo i coati, una specie della famiglia dei Procionidi, di cui aveva visto alcuni esemplari durante il suo viaggio, aggiungendo nell'articolo relativo che l'animale è "*conosciuto anche come l'oritteropo Brasiliano*"[4], senza ovviamente citare nessuna fonte in quanto l'informazione era inventata da lui, conseguenza del fatto che aveva scambiato gli esemplari di coati per oritteropi, mammiferi di diversa specie appartenenti alla famiglia degli Orycteropodidae. Un'informazione di questo tipo non è catalogabile come vandalismo, in quanto non si tratta di una nozione volutamente errata, ma secondo le linee guida di Wikipedia di cui si è parlato nella precedente sezione, richiede-

[3]Wikipedia: Edit filter

[4]Brazilian aardvark

rebbe quantomeno di essere marcata come *citation needed* in quanto si tratta di un'informazione aggiuntiva di cui bisognerebbe provare la provenienza da una fonte attendibile. In realtà, questa informazione non è stata segnalata, per cui è rimasta sulla pagina per oltre un anno. Nel corso del tempo però il termine *oritteropo Brasiliano* non solo è rimasto presente su Wikipedia, ma è stato anche adottato da diversi siti e testate online per riferirsi ai coati: sono presenti riferimenti sull'Indipendent [5], il Daily Mail[6], addirittura un libro pubblicato dall'Università di Chicago[7] [32]. Questo mostra come nonostante l'esistenza di linee guida e gli inviti all'utilizzo delle fonti, vi siano casi in cui l'assenza di fonti per le informazioni aggiunte su una pagina non venga segnalata come *citation needed* anche dopo molto tempo l'inserimento della stessa; ma il problema riguardo questo specifico caso va oltre, in quanto all'informazione inserita dal ragazzo nella pagina di Wikipedia nel 2010 era stata aggiunto un riferimento bibliografico ad un articolo del Telegraph [8], creando così un caso di circular reporting, comportamento descritto anche in una pagina di Wikipedia stessa [9] e con il quale l'enciclopedia libera online ha avuto spesso a che fare per le pagine relative a personaggi famosi come l'attore Sacha Baron Cohen [32]. Ad oggi, la pagina di Wikipedia sull'orit-

[5] From wallabies to chipmunks, the exotic creatures thriving in the UK, http://www.independent.co.uk/environment/nature/from-wallabies-to-chipmunks-the-exotic-creatures-thriving-in-the-uk-2006096.html

[6] Hunt for the runaway aardvark: Lady McAlpine calls on public to help find her lost ring-tailed coat, ihttp://www.dailymail.co.uk/news/article-2305602/Hunt-runaway-aardvark-Lady-McAlpine-alls-public-help-lost-ring-tailed-coati.html

[7] TheBookofBarelyImaginedBeings:A21stCenturyBestiary,CasparHenderson, UniversityofChicagoPress2013

[8] Scorpions, Brazilian aardvarks and wallabies all found living wild in UK, study finds. http://www.telegraph.co.uk/earth/wildlife/7841796/Scorpions-Brazilian-aardvarks-and-wallabies-all-found-living-wild-in-UK-study-finds.html

[9] Circular reporting: https://en.wikipedia.org/wiki/Circular_reporting

teropo Brasiliano [10] è ancora presente e contiene il redirect automatico alla pagina sui Coati [11]. Il caso dell'oritteropo Brasiliano è un esempio che mostra non solo il rischio di persistenza sulle pagine di informazioni errate, ma addirittura che queste informazioni possono essere utilizzate come fonte per la creazione di una verità sbagliata ma accettata dalla massa. Questi casi sono dovuti al fatto che le informazioni usate dai giornalisti non sono state verificate prima di essere state utilizzate, causando la diffusione dell'informazione errata e contribuendo al fenomeno di *Wikiality* definito da Stephen Colbert: "chiunque può modificare una pagina e se un numero sufficiente di utenti è d'accordo con la modifica, questa diventa verità", come è accaduto nel caso della giorno di nascita di Jimmy Wales, per il quale Wikipedia e la massa hanno dato il via ad un circolo vizioso che ha per così dire "imposto" che la data fosse il 7 Agosto, nonostante Wales stesso affermi che la data sia in realtà sbagliata a causa di un errore sul certificato di nascita originale utilizzato come fonte per l'informazione.

1.3.2 Autocorrezione e resilienza

In letteratura sono presenti diversi casi di studi ed esperimenti riguardo come e quanto persistano errori inseriti volontariamente all'interno di articoli di Wikipedia. Come riportato da P.D. Magnus in un esperimento del 2008, il giornalista Dan Tynan ha inserito informazioni fittizie all'interno di pagine riguardanti biografie di importanti personaggi del mondo del mercato delle tecnologie, verificando che gli errori sono statti rimossi soltanto 3 mesi e 200 modifiche dopo[10]. Ma altri esperimenti, anche questi senza l'uso di approcci più scientifici, mostrano come invece i contributi erronei possano essere corretti molto in fretta. Questo è il caso dell'esperimento riportato da Halavais del 2006, durante il quale sono state inserite 13 informazioni errate utilizzando uno pseudonimo. Tutte le informazioni sono state rimosse entro

[10] `https://en.wikipedia.org/w/index.php?title=Brazilian_aardvark&redirect=no`

[11] `https://en.wikipedia.org/wiki/Coati#Nickname_misconception`

le prime 3 ore. Quest'ultimo caso però rappresenta un esempio specifico di tecnica di difesa di Wikipedia, in quanto la rimozione di tutti gli errori è stato dovuto ad un *association effect* che ha permesso agli utenti di capire che lo pseudonimo dell'esperimento di Halavais fosse in realtà un utente malevole e che quindi tutti i suoi contributi erano catalogabili come vandalismi e per cui dovevano essere rimossi in blocco. Magnus stesso, in un suo studio del 2008[11], ha effettuato un esperimento di questo tipo, inserendo 36 notizie fittizie all'intero di altrettante pagine bibliografiche di filosofi (quindi non inerenti concetti filosofici), uniformando in qualche modo la categoria degli articoli oggetto dell'esperimento, presupponendo fossero mantenute da comunità di utenti similari. Le frasi inserite non includevano link e sono state inserite in punti plausibili della pagina, alcune menzionavano fonti ma non vi erano citazioni. Di seguito due esempi di frasi riportate dallo studio di Magnus, rispettivamente riguardanti Boethius e Gilbert Ryle:

'It is known that he lost two fingers on his left hand in a childhood accident, although there is no record of how exactly it occured.'

'After retiring, Ryle bought a small farm. He tinkered with automated processes to care for livestock, although they never proved to be commercially viable.'

Le frasi sono state inserite in 12 gruppi di 3 frasi per volta per evitare l'*association effect* e da IP diversi per ogni gruppo di inserimento (12 gruppi da 3). Questi accorgimenti hanno limitato ma non rimosso il problema dell'*association effect*. Le pagine oggetto delle modifiche erano accettate come valide da Wikipedia, alcune *in vetrina* (e di conseguenza più controllate), e sono state monitorate ognuna per 48 ore. Dall'esperimento è emerso che 15 frasi sono state rimosse nelle 48 ore e 3 sono state segnalate come *citation needed*. Per cui circa 1/3 degli errori corretti entro le 48 ore e 1/4 hanno avuto *assocation effect*. Magnus afferma che ingrandendo l'esperimento con più errori avrebbe ad un maggiore richiamo di attenzione da parte degli autori di Wikipedia per cui l'esperimento non sarebbe stato valido.

Secondo alcuni studi, come mostrato nello da Priedhorsky [8], l'insieme

delle tecniche di difesa fanno sì che Wikipedia sia effettivamente un sistema affidabile e resiliente. Queste conclusioni derivano da esperimenti che mostrano che il 42% degli errori viene riparato entro una visualizzazione dell'errore stesso, che coincide con la persistenza media degli errori di 2.8 minuti sulla pagina rilevata da Viégas [12]. Dei restanti errori, l'11% ha superato le 100 visualizzazione e solo lo 0,06% le 1000 visualizzazioni, portando alla probabilità del 0,0037 di incontrare una pagina danneggiata nel 2003, valore che raggiunge il 0,0076 nel 2006. La persistenza rimane stabile nel tempo, quindi l'aumento di probabilità di visualizzare un danno è dovuta ad una maggior frequenza dei danni. A partire dal 2006 sono stati introdotti dei *bot* di auto-riparazione che hanno arrestato la crescita esponenziale della probabilità di incontrare un danno minore. I bot in quanto sistemi automatizzati difficilmente possono correggere errori di carattere semantico, limitandosi alle problematiche di forma e grammatica.

Anche nei casi in cui gli errori arrivassero agli utenti e fossero riconoscibili come tali, questi non costruirebbero nella loro mente un'idea negativa riguardo al soggetto dell'articolo contenente un errore, quanto piuttosto un'idea di inaffidabilità di Wikipedia; ciò mostra ancora come gli utenti formano delle idee in base a come utilizzano Wikipedia.

1.3.3 Resilienza di Wikipedia in termini di rete

Le capacità di resilienza di Wikipedia saranno differenti a seconda che se ne prenda in analisi la rete strutturale o la rete socio-produttiva. Nel primo caso siamo di fronte ad una resilienza strutturale, tipica del Web, grazie alla quale Wikipedia è protetta dalla cancellazione malevole di collegamenti tra le pagine poichè, data la struttura di rete, sarà possibile comunque raggiungere la pagina attraverso altri collegamenti [14, 7]. I dettagli di questi comportamenti non sono analizzati in questo studio poichè ci si è focalizzati esclusivamente sull'aspetto semantico della resilienza del sistema.

Per quanto riguarda invece la rete socio-produttiva, la resilienza in questa struttura comporta una duplice conseguenza: un attacco di tipo semantico

all'interno di una pagina del core porta ad un alta probabilità di correzione dovuta al fatto che molti collegamenti permettono di raggiungere la pagina, ma allo stesso tempo l'*infezione* della pagina può causare una maggior diffusione dell'errore.

Tramite un approccio empirico rappresentato dall'iniezione di errori all'interno delle pagine selezionate come campione e dalla misurazione dei tempi e delle modalità di correzione di questi, sarà possibile studiare le capacità di autocorrezione semantica di Wikipedia. Ci si aspetta l'esistenza di un circuito di autoregolazione all'interno del sistema, che porti alla correzione degli errori inseriti con tempistiche e modalità differenti a seconda delle variabili: tipologia e gravità dell'errore, tipo e categoria della pagina. Si suppone inoltre, come ipotesi più avanzata, che il processo di autocorrezione possa essere influenzato anche dalla metodologia di iniezione degli errori e dal tipo di utenti che andranno a correggere gli errori, per cui sarà necessaria un'ulteriore analisi anche riguardo l'approccio di utilizzato nell'inserimento degli errori così da costruire esperimenti differenti. Le analisi per la selezione degli errori, delle variabili e della metodologia di iniezione degli errori saranno descritte nel Capitolo 3.

Capitolo 2

Autocorrezione e resilienza semantica

2.1 Modello teorico di studio delle proprietà autocorrettive

Per rappresentare Wikipedia ed i processi sottostanti, si è ipotizzato un modello composto dai seguenti 3 elementi principali:

- Prodotto
- Processo
- Struttura

Il prodotto è Wikipedia inteso come oggetto di fruizione passiva, l'enciclopedia gratuita online e liberamente modificabile. La sua caratteristica è quella di essere **visibile**.
Il processo rappresenta l'insieme delle modifiche, delle revisioni e quindi della cronologia legate ad una pagina di Wikipedia. La caratteristica è quella di essere **rintracciabile** e, diversamente dal prodotto, è presente nel modello in un numero elevato di entità, ciascuna corrispondente a ogni processo di

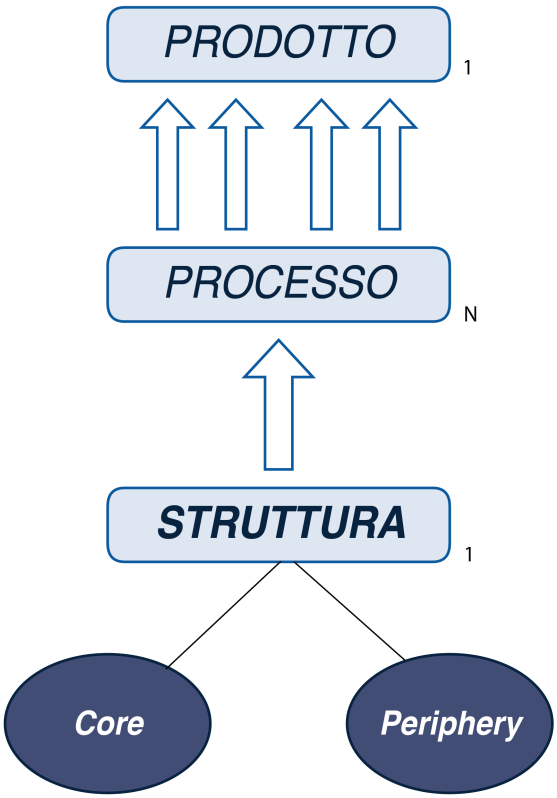


Figura 2.1: Modello strutturale

creazione o modifica di una pagina. Gli elementi che compongono il processo sono il **tempo** e gli **attori**.

La struttura, infine, anch'essa presente in un'unica entità all'interno del modello, è rappresentata dalla comunità dei soggetti e costituisce l'oggetto dello studio.

Rappresentato tramite questo modello quindi, Wikipedia è un prodotto fruibile, che nasce dall'unione di un numero di processi sottostanti, a loro volta generati da una struttura che rappresenta la community. All'interno di questo modello la nostra attenzione è focalizzata sulla struttura, oggetto principale della prima parte dello studio, per la quale (figura 2.2) si ipotizza l'utilità a fini sperimentali di distinguere fra pagine di natura "generale"

(ovvero dotate, in ragione dei loro contenuti, di una elevata probabilità di accesso da parte di un ampio insieme di utenti) e pagine "specifiche", in quanto relative a contenuti di dominio più ristretto, a cui corrisponde una minor numerosità degli utenti potenzialmente in grado di rilevare un errore e, a maggior ragione, di assumere un ruolo attivo nella sua gestione. Si tratta di una distinzione preliminare di natura qualitativa, di indirizzo per la successiva caratterizzazione topologico-quantitativa delle effettive pagine oggetto di esperimento.

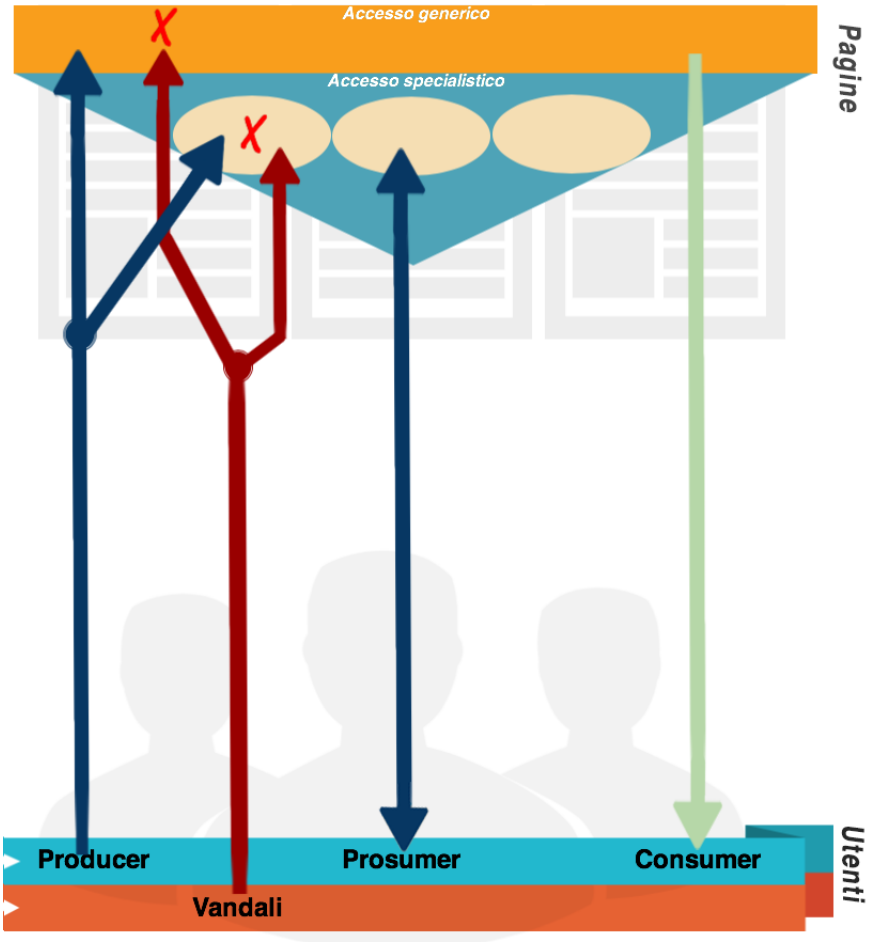


Figura 2.2: Modello teorico di interazione tra i livelli di rete: pagine e utenti, che rispecchia le proprietà autocorrettive

Per quanto riguarda gli utenti, si ipotizza una suddivisione nelle seguenti categorie:

Producer: Sono utenti che possiedono ruoli attivi nella gestione di Wikipedia producendo contenuto di qualità ma anche applicando modifiche sulla forma e sulla struttura delle pagine. Relativamente all'accesso alle pagine, scrivono in maniera diffusa e trasversale indipendentemente dall'argomento di riferimento.

Prosumer: fusione del termine Producer e Consumer, si riferisce a utenti registrati che producono contenuto limitatamente agli argomenti per i quali sono esperti, scrivendo quindi solo su pagine appartenenti a determinate categorie. Questo tipo di utente è il generico Wikipediano, che fruisce di Wikipedia in lettura ma produce anche contenuto di qualità su argomenti di sua conoscenza.

Consumer: utenti visitatori che accedono a Wikipedia anche in maniera anonima e fruiscono dell'enciclopedia in termini di semplice consultazione. Non contribuiscono alla produzione del contenuto, se non con modifiche minori e comunque trascurabili dall'analisi in quanto non tracciabili poichè eseguite senza registrazione al sito.

Vandali: non costituisce una vera e propria categoria di utilizzatori di Wikipedia, ma rappresenta gli utenti malevoli che inseriscono gli errori nelle pagine durante gli esperimenti dello studio scrivendo su pagine di entrambe le tipologie.

Per tutte le categorie di utenti, l'accesso in lettura è lo stesso ed equivalente per i due schemi di accesso alle pagine, per cui nel modello non è stata rappresentata questa tipologia di interazione, se non per gli utenti Consumer.

Si ipotizza quindi la seguente dinamica per la correzione degli errori a seconda della classe di utenti:

- Producer: correzione di errori inseriti sia in pagine generiche che in pagine specialistiche effettuata tendenzialmente attraverso *revert* che ri-

portano la pagina alla versione cronologicamente precedente, in quanto i Producer non sono esperti di ogni categoria ma lavorano trasversalmente su tutte le categorie.

- Prosumer: correzione di errori che si trovano su pagine specialistiche. La gestione dell'errore potrà avvenire sia tramite una *revert* ma anche con una correzione dettagliata.

2.2 Sviluppo del modello di rete

Le strutture di rete ricavate dall'analisi dei dati rappresentano i criteri per la selezione delle pagine oggetto dell'iniezione degli errori nella seconda parte dello studio che sarà descritta nella prossima sezione, costituendo un metodo per stimare le diverse probabilità di autocorrezione per ciascuna pagina. Queste probabilità si ipotizzano influenzate sia dalle categorie di appartenenza delle pagine che dalla tipologia di queste, variabili che saranno evidenziate nelle reti.

2.2.1 Selezione del campione

Per limiti di calcolo non è stato possibile utilizzare l'intero insieme dei dati disponibile, si è scelto quindi di estrarre un campione stratificato ed eterogeneo in modo da rappresentare il più possibile l'universo della versione in lingua italiana di Wikipedia. Dal dump XML contenente tutte le modifiche a tutte pagine (da qui identificate con il termine *revision*) dal 2003 ad oggi, si è deciso di estrarre un campione di pagine appartenente a 3 categorie, descritte di seguito. Nelle tabelle 2.1, 2.2 e 2.3 è riportato l'elenco delle pagine selezionate per ogni gruppo. Anche all'interno di ogni categoria, per limiti di calcolo è stato necessario stringere il campione ad un numero limitato di pagine appartenenti a ciascuna categoria, in quanto si è scelto di non applicare un limite temporale ma di tenere conto di tutte le revision presenti su queste

pagine dalla data di creazione della pagina fino alla data di esecuzione dello studio, ovvero Maggio 2014.

Gruppo 1: Pagine ad accesso indifferenziato Le pagine selezionate appartengono ad una categoria caratterizzata dal fatto che gli utenti che apportano modifiche alle pagine possono essere sia utenti specializzati nell'argomento ma anche, e soprattutto, utenti generici senza nessuna specializzazione a riguardo, che accedono spesso in modifica su queste pagine. Questa caratteristica è dovuta al fatto che l'argomento trattato da queste pagine è di competenza comune per il quale un alto numero di persone può contribuire fornendo la sua conoscenza e opinione. La categoria selezionata per questo gruppo è il *calcio*.

Gruppo 1: **Calcio**	
Titolo pagina	**Nr. revision**
Arsenal Football Club	1447
Associazione Calcio Milan	2101
Campionato mondiale di calcio	1388
Edinson Cavani	1836
Football Club Internazionale Milano	8367
Juventus Football Club	7922
Marcatori dei campionati italiani di calcio	657
Serie A	2737
Statistiche della Serie A	796
UEFA Champions League	1764

Tabella 2.1: Nr di revision da parte di utenti distinti per ogni pagina per gruppo Calcio

Gruppo 2: Pagine ad accesso specializzato ma influenzato dal trend Le pagine appartenenti a questa categoria sono generalmente modificate da pochi utenti, specializzati sull'argomento. A causa però di un

incremento dell'interesse pubblico sull'argomento dovuto a fatti di comune interesse sociale, queste pagine subiscono una variazione nel loro classico schema di accesso e di modifica, per cui in un certo periodo temporale ricevono più modifiche rispetto al solito, anche da parte di utenti che non sono necessariamente quelli specializzati che solitamente le modificano. Queste pagine nel nostro studio sono accomunate dall'essere relative ad argomenti riguardanti la *Crisi in Ucraina del 2014*.

Gruppo 2: **Crisi in Ucraina del 2014**	
Titolo pagina	**Nr. revision**
Crisi della Crimea del 2014	23
Elezioni presidenziali ucraine del 2010	25
Euromaidan	21
Julija Tymosenko	88
Kiev	172
Primi ministri dell'Ucraina	17
Repubblica autonoma di Crimea	155
Russia	408
Trattato di adesione della Crimea alla Russia	4
Ucraina	321

Tabella 2.2: Nr di revision da parte di utenti distinti per ogni pagina per gruppo Crisi in Ucraina del 2014

Gruppo 3: Pagine ad accesso altamente specializzato Queste pagine appartengono ad una categoria specialistica, di nicchia, le cui modifiche vengono tendenzialmente fatte da pochi utenti altamente specializzati sull'argomento. Nello studio si è scelto di estrarre pagine legate al Neorealismo italiano.

Date le differenti caratteristiche delle categorie, ci si aspetta che la struttura sottostante ad ognuna di queste si differenzi dall'altra e questo influenzi

Gruppo 3: **Neorealismo italiano**			
Titolo pagina	**Nr. revision**	**Titolo pagina**	**Nr. revision**
Aldo Fabrizi	125	Neorealismo (cinema)	65
Anna Magnani	175	Nouvelle Vague	70
Cinema (rivista)	13	Nová vlna	7
Cinéma vérité	1	Nuovo cinema tedesco	18
Festival del cinema neorealistico	2	Ossessione (film 1943)	54
La terra trema	51	Realismo poetico	19
Ladri di biciclette	114	Riso amaro	59
Luchino Visconti	169	Roberto Rossellini	134
Michelangelo Antonioni	143	Roma città aperta	86
Neorealismo	74	Vittorio De Sica	180

Tabella 2.3: Nr di revision da parte di utenti distinti per ogni pagina per gruppo Neorealismo

il numero e la tipologia di modifiche alle pagine. Per il gruppo di pagine ad accesso indifferenziato, ci si aspetta un alto numero di revision effettuate da un alto numero di utenti distinti, che contribuiscono, anche con piccole modifiche, all'evoluzione delle pagine. Diversamente invece per le altre due categorie, che presenteranno un numero di revision più limitato, effettuato da un numero di utenti più basso. Con queste ipotesi di partenza sul campione, confermate da una analisi statistica basata sul numero di revision di alcune pagine campione per ciascuna categoria[1], si è deciso di estrarre un numero di pagine minore per quanto riguarda il gruppo 1 rispetto agli altri due gruppi, poichè con già un esiguo numero di pagine si ricaverebbe lo stesso numero di revision estratte dai gruppi 2 e 3. Allo stesso tempo però è stato necessario

[1]Fonte statistiche per pagina: `http://vs.aka-online.de/cgi-bin/wppagehiststat.pl`

mantenere un certo livello di eterogeneità tra il numero di pagine del gruppo 1, selezionando quindi almeno 10 pagine che rappresentassero diversi ambiti della stessa categoria (squadre di calcio, campionati, classifiche, giocatori) e di conseguenza la mole di dati di partenza di questo gruppo risulta molto maggiore di quella degli altri gruppi. Inoltre, per quanto riguarda il gruppo 3, è stato necessario raccogliere dati da un numero di pagine doppio di quello raccolto con gli altri due gruppi per poter raggiungere un numero di revision sufficiente per l'analisi. Nello specifico il numero di revision per ogni gruppo è rappresentato nella tabella 2.4

Gruppo	**Nr. pagine**	**Nr. revision**	**Anno prima revision**	**Anno ultima revision**
Calcio	10	29015	2003	2014
Crisi Ucraina	10	3358	2004	2014
Neorealismo	20	3358	2004	2014

Tabella 2.4: Numero di pagine e revision per Gruppo

2.2.2 Metodologia di raccolta del campione

Pagine

Il seguente blocco di codice estratto dal dump XML riporta un esempio di due revision per la pagina "Neorealismo (cinema)" :

```
<page>
    <title>Neorealismo (cinema)</title>
    <ns>0</ns>
    <id>64438</id>
    <revision>
      <id>329872</id>
```

```
      <timestamp>2004-12-12T12:51:38Z</timestamp>
      <contributor>
        <username>Shaka</username>
        <id>2225</id>
      </contributor>
      <comment>da finire di tradurre</comment>
      <text id="329872" bytes="7158" />
      <sha1>3re4vj1x3cj7n20ocgm2dprnt26amx9</sha1>
      <model>wikitext</model>
      <format>text/x-wiki</format>
    </revision>
    <revision>
      <id>334496</id>
      <parentid>329872</parentid>
      <timestamp>2004-12-21T23:47:58Z</timestamp>
      <contributor>
        <username>Franztanz</username>
        <id>2788</id>
      </contributor>
      <comment>un po' avanti ...</comment>
      <text id="334496" bytes="7274" />
      <sha1>dg029k69ohvdk5dosbzj8qiv8fef5bq</sha1>
      <model>wikitext</model>
      <format>text/x-wiki</format>
    </revision>
  ...
</page>
```

I dati di interesse per lo studio sono, per ciascuna delle pagine selezionate:

- `title`, il titolo della pagina
- `revision id`, identificativo univoco della revision

- `timestamp`, data della revision
- `username`, nome dell'utente che ha fatto la revision
- `username id`, identificativo univoco dell'utente

Sebbene Wikipedia metta a disposizione strumenti per l'estrazione delle informazioni dalle pagine[2], questi strumenti non si sono rivelati sufficienti per ricavare i dati necessari allo studio, a causa di limitazioni sul numero di revision esportabili per ogni pagina; per ricavare quindi i dati riguardanti le pagine selezionate, è stato necessario utilizzare come sorgente il dump XML. Per poter gestire la manipolazione di un file XML di grandi dimensioni come il dump (circa 30GB) è stato necessario implementare un parser Java che prendesse in input il file e le pagine per le quali estrarre le informazioni, e restituisse i dati come output in formato CSV. Il parser è basato sulle librerie SAX[3] per la gestione dei file XML in quanto possiede una gestione della memoria più leggera rispetto alle librerie DOM, e permette di analizzare l'intero file e produrre l'output in pochi minuti. É stato prodotto un file CSV per ognuno dei 3 gruppi di pagine e i dati sono stati caricati su corrispondenti strutture tabellari in un database MySQL locale. I listati del parser Java sono disponibili in Appendice A.

Utenti

Dal dump sono quindi state estratte tutte le informazioni riguardanti le revision, compreso l'id che identifica l'utente che ha effettuato la modifica alla pagina. All'interno di questo dump non sono però presenti ulteriori informazioni riguardo gli utenti, perciò per ricavare informazioni dettagliate sugli utenti come ad esempio i gruppi ed i ruoli di appartenenza, è stato necessario utilizzare i servizi di query esposti da MediaWiki tramite le API di Wikipedia[4]. Per estrarre le informazioni in blocco, sono stati sviluppati

[2]Pagina speciale per l'esportazione in formato XML delle pagine MediaWiki, http://it.wikipedia.org/w/index.php?title=Speciale:Esporta

[3]http://www.saxproject.org

[4]MediaWiki API, https://www.mediawiki.org/wiki/API:Query

dei semplici script bash per lanciare i comandi cURL[5] per l'esecuzione delle query verso Wikipedia. Per ciascun utente ricavato dall'elenco delle revision I parametri di query delle API utilizzati per recuperare le informazioni sugli utenti sono stati:

- `format=xml`, per ricavare l'output in formato XML in modo da poterlo successivamente manipolare più agevolmente.
- `ususers=`, per specificare il nome dell'utente di cui si vogliono ricavare i dettagli.
- `usprop=blockinfo|groups|editcount|registration`, per specificare quali informazioni estrarre: se l'utente è bloccato, i gruppi di appartenenza, il numero di modifiche effettuate e la data di registrazione.

Di seguito un esempio di query effettuata per un utente:

```
curl 'http://it.wikipedia.org/w/api.php?action=query
     &format=xml&list=users&ususers=Alberteinstein
     &usprop=blockinfo|groups|editcount|registration'
```

Ogni query ha prodotto un file XML da cui sono state estratte le informazioni necessarie all'analisi attraverso un parser Java utilizzando le librerie DOM. I dati estratti sono stati caricati su apposite tabelle sullo stesso database MySQL locale in modo da poter essere manipolati insieme ai dati relativi alle pagine estratti dal dump XML. I listati dei programmi sono disponibili in Appendice A.

Il numero utenti distinti per ciascuna categoria di pagine per i quali è stato necessario estrarre le informazioni è riportato in tabella 2.5.

Come descritto all'interno di Wikipedia in apposite pagine tecniche [42, 43], gli utenti registrati sul portale possono appartenere a gruppi che gli forniscono delle funzionalità che gli utenti normali non possiedono. Tutti gli utenti registrati appartengono al gruppo *user*, ma poichè la politica di Wikipedia è quella della completa libertà di accesso e modifica [44], è possibile

[5] cURL e libcurl, http://curl.haxx.se

Gruppo	Nr. utenti distinti
Calcio	2855
Crisi Ucraina	760
Neorealismo	729

Tabella 2.5: Numero di utenti distinti che hanno modificato pagine del gruppo

che alcune modifiche siano state effettuate da utenti non registrati, "utenti anonimi" identificati tramite l'IP con cui accedono al sito, che possono creare e modificare pagine, non possono spostarle e non possono caricare file; sono inoltre obbligati a effettuare un CAPTCHA quando devono inserire un collegamento esterno. Questi utenti sono stati necessariamente esclusi dallo studio in quanto una modifica da parte di un utente non registrato non è associabile ad uno specifico soggetto all'interno della community. L'esclusione di questi soggetti è stata effettuata già al momento del parsing del dump XML, in cui le revision prodotte da IP non sono state prese in considerazione. Gli utenti registrati invece possono appartenere a diversi gruppi:

autoconfirmed Sono utenti registrati da più di 4 giorni che effettuano il login; l'attribuzione del flag è automatica e dà la possibilità di: caricare file o nuove versioni di file, spostare pagine, modificare pagine protette parzialmente, verificare le modifiche di altri utenti, salvare collegamenti esterni senza bisogno di inserire il CAPTCHA

sysop Sono utenti "Amministratori", volontari che hanno avuto la fiducia dalla comunità degli utenti per poter compiere determinate azioni tecniche[6]

checkuser Amministratori autorizzati a controllare quali IP corrispondono a un dato nome utente e viceversa. L'uso più comune di questa funzione è quello di verificare i *sockpuppet* degli utenti bloccati.

[6]http://it.wikipedia.org/wiki/Wikipedia:Amministratori

bureaucrat Amministratori che hanno avuto la fiducia dalla comunità degli utenti per poter compiere determinate azioni tecniche necessarie per il corretto funzionamento di Wikipedia

autopatrolled Sono utenti detti "autoverificati", utenti registrati i cui contributi sono verificati in automatico. Serve per ridurre il numero delle modifiche segnalate come *da verificare*, incrementando pertanto la possibilità di rimuovere modifiche effettivamente dannose o improprie. L'attività di verifica dei contributi altrui non comporta infatti una valutazione di merito sul contributo verificato, ma attesta solo la presumibile assenza di palesi vandalismi o di modifiche a vario titolo improprie. Un utente è quindi adatto ad essere proposto per il gruppo degli autoverificati non perchè la sua contribuzione sia esente da errori, ma perchè la sua affidabilità appare sufficiente perchè i patroller non debbano sorvegliare ogni sua modifica

filemover Sono wikipediani registrati e autoverificati che hanno la possibilità di rinominare i file spostandoli a nomi idonei

rollbacker Utenti registrati dediti assiduamente al patrolling che hanno ricevuto un'abilitazione aggiuntiva, il tasto rollback, grazie al quale possono operare più velocemente nel ripristino in tempo reale delle pagine danneggiate da vandalismi, inserimenti erronei o rimozioni totali del contenuto di voci e pagine di servizio.

bot Programmi automatici che lavorano sulle pagine

ipblock-exempt Un utente esente dal blocco IP è un utente registrato sul quale non agiscono i blocchi di range di IP. gli utenti che invece non appartengono al gruppo degli amministratori possono richiedere un'esenzione dal blocco IP solo se sono in grado di mostrare delle motivazioni valide, credibili e inconfutabili.

accountcreator Un creatore di utenze è un utente dotato di funzioni particolari per la creazione di utenze.

In base al gruppo di appartenenza, sono stati definite 3 tipologie di utente per semplificare la loro rappresentazione nell'analisi:

Utenti normali , utenti che possiedono il gruppo *autoconfirmed*. All'interno di questa tipologia sono stati inseriti anche gli utenti appartenenti al gruppo *accountcreator* in quanto sono effettivamente utenti normali con alcuni diritti aggiuntivi che non sono rilevanti per lo studio e sono raramente usati.

Utenti supervisori , utenti appartenenti ai gruppi *autopatrolled*, *filemover* o *rollbacker*, i quali svolgono attività di patrolling

Utenti amministratori , utenti appartenenti ai gruppi *sysop*, *bureacrat*, *checkuser*, che possiedono diritti gestionali e sono autorizzati ad attività tecniche su Wikipedia.

La distribuzione degli utenti per ciascuna tipologia è riportata in figura 2.6.

Categoria	**Nr. utenti**
Amministratori	107
Supervisori	473
Normali	3785
BOT	210
TOTALE	4574

Tabella 2.6: Numero di utenti per ogni tipologia

Questi gruppi sono stati associati alle categorie di utenti definite nel modello rappresentato in figura 2.2:

- Producer= Utenti supervisori e Utenti amministratori
- Prosumer=Utenti normali
- Consumer, nessuna mappatura con i gruppi in quanto gli utenti non registrati non sono stati presi in considerazione nell'analisi.

Oltre gli utenti anonimi, anche i *bot* sono stati esclusi dall'analisi, poichè non sono utenti umani ma programmi sviluppati da utenti per svolgere funzionalità macchinose, principalmente di correzione ortografica, i cui contributi quindi non si possono ritenere rilevanti ai fini dello studio e anzi rappresenterebbero un elevato numero di contributi per cui influenzerebbero notevolmente i risultati. Per escludere questi utenti, non è bastato identificare gli utenti che risultavano appartenere al gruppo bot, ma anche quelli le cui ultime lettere dell'username coincidevano con la stringa *bot*. Questo è dovuto al fatto che se un bot è inattivo, è stato rimosso dal gruppo bot ma quando era attivo ha comunque eseguito delle modifiche.

2.2.3 Analisi del campione

I dati risultanti dalla combinazione delle informazioni estratte dal dump e quelle dei singoli utenti caricati sul database MySQL locale sono stati interrogati tramite specifiche query per estrarre le informazioni di interesse. Questi risultati mostrano la presenza di distribuzioni power law all'interno del sistema. A seguire questa distribuzione di frequenza sono:

- il numero di modifiche alle pagine da parte degli utenti. Si evidenzia l'invarianza di scala, poichè la stessa proprietà vale sia per ciascuna categoria (si vedano le tabelle e le figure 2.18, 2.20, 2.22) che per il totale delle pagine, come mostrato in nella tabella e nel grafico relativi 2.16)

- il numero di pagine modificate in comune per coppie di utenti (tabelle e figure 2.8, 2.12, 2.10, 2.14)

- il numero di utenti in comune per coppie di pagine (tabella e figura 2.7). I dati si riferiscono a un intervallo temporale che comprende unicamente gli anni dal 2011 al 2014.

Per quanto riguarda le modifiche sulle pagine, si nota la corrispondenza dei risultati ricavati con quelli evidenziati dagli studi in letteratura riportati

nel Capitolo 1: pochi utenti che producono la maggior parte delle modifiche e molti utenti che scrivono poco. Questo comportamento appare sia sul totale delle modifiche per tutti i set sperimentali selezionati, che per ogni singolo set.

2.3 Risultati di rete

Wikipedia è caratterizzabile attraverso Network Analysis sia dal punto di vista della struttura di prodotto (le relazioni fra le pagine), sia dal punto di vista della struttura del soggiacente sistema socio-produttivo (le relazioni fra gli user dotati di ruoli attivi di redazione, monitoraggio ed amministrazione). Dall'analisi dei dati estratti dal campione attraverso i software di analisi delle reti[1], emerge l'evidenza di una strutture di rete che segue un modello di tipo core-periphery; questa rispecchia la distribuzione power law dei dati: il *core* è costituito dalla coda destra del grafico, ovvero pochi nodi con un alto valore di legame (i valori sulle ascisse). Questa topologia di rete si evidenzia sia per quanto riguarda le relazioni tra pagine che quelle tra attori, come sarà descritto dettagliatamente nelle prossime sezioni.

La tecnica più efficiente di analisi richiede la preliminare costruzione di un dataset bimodale (detto "attori/eventi")[5], costituito da una matrice m x n, dove m è il numero di pagine assunte a priori nel campione (gli eventi) ed n è il numero degli utenti singoli che, dall'inizio della storia della pagina al momento del campionamento hanno agito in scrittura su una o più pagine (gli attori). La valorizzazione delle singole celle della matrice esprime il numero di volte in cui un utente ha acceduto in editing alla relativa pagina.

Attraverso un procedimento di affiliazione, il dataset bimodale è ridotto alternativamente a due dataset monomodali, relativi rispettivamente alle relazioni che gli attori hanno fra loro attraverso le pagine condivise (matrice n x n) ed alle relazioni che le pagine hanno fra loro attraverso gli attori redazionalmente comuni (matrice m x m), ovvero che hanno lavorato, anche in tempi diversi, su di esse. Entrambi i dataset sono una *proxy* del sistema di produzione:

- il primo è rivolto a studiare l'articolazione dei ruoli emergenti, con particolare attenzione al rapporto fra gli user di tipo "Prosumer" e

[1] Ucinet 6.1; Netdraw 2.1

quelli più nettamente "Producer", in essi inclusi quelli dotati di ruoli formali (amministratori, ...);

- il secondo è rivolto a studiare l'articolazione dei rapporti fra pagine, non sulla base dei link fra di esse esistenti, ma del "tessuto connettivo" dato dal sistema socio-produttivo soggiacente. Pagine significativamente distanti per contenuto possono in realtà essere alla istanza di un solo passo, ove abbiano uno o più attori operati in entrambe.

Rispetto al disegno dell'esperimento, entrambi i dataset sono particolarmente utili per affinare le ipotesi e definire le logiche di campionamento. Possono infatti essere immaginati due meccanismi principali di avvistamento e correzione degli errori introdotti intenzionalmente da atti vandalici:

- uno di tipo "comunitario", dove la regolazione è svolta dai "Prosumer" aggregati attorno alle pagine di dominio di loro interesse;
- un secondo di tipo "istituzionale" dove la regolazione è svolta dai "Producer", non necessariamente sulla base della comprensione delle effettive caratteristiche dell'errore (visto che essi non possono essere esperti in tutti i domìni di contenuto), ma attraverso applicazione di regole generali e/o di funzioni di tracking del comportamento dei supposti vandali.

In ipotesi, i due meccanismi sono compresenti e fra loro competitivi, sulla base delle caratteristiche dei network monomodali sopra introdotti. Dal punto di vista analitico, l'esame dei due dataset ottenuti per affiliazione mette in evidenza alcune caratteristiche di struttura della porzione di Wikipedia oggetto di esperimento.

É del tutto rilevante osservare che - pur a fronte di una forte distanza simbolico-percettiva fra i tre domìni presi in esame (Calcio, Neorealismo e Crisi in Ucraina del 2014) - entrambi i network monomodali si presentano caratterizzati da una macro-componente comune. Appare dunque respinta la possibile ipotesi che legge Wikipedia come un insieme di comunità di dominio; esistono invece significative "legature" fra i diversi cluster di pagine

tematiche. Il che potrebbe costituire una maggior garanzia di autocorrezione, almeno attraverso il meccanismo del tracking del comportamento del supposto vandalo.

2.3.1 Rete tra utenti

La prima osservazione importante riferita alla rete "attori/attori" relativa all'intero insieme delle pagine pre-campionate, acquisibile anche visivamente (oltreché verificata in via statistica), è la presenza di una struttura core-periphery, come osservabile in figura 2.3.

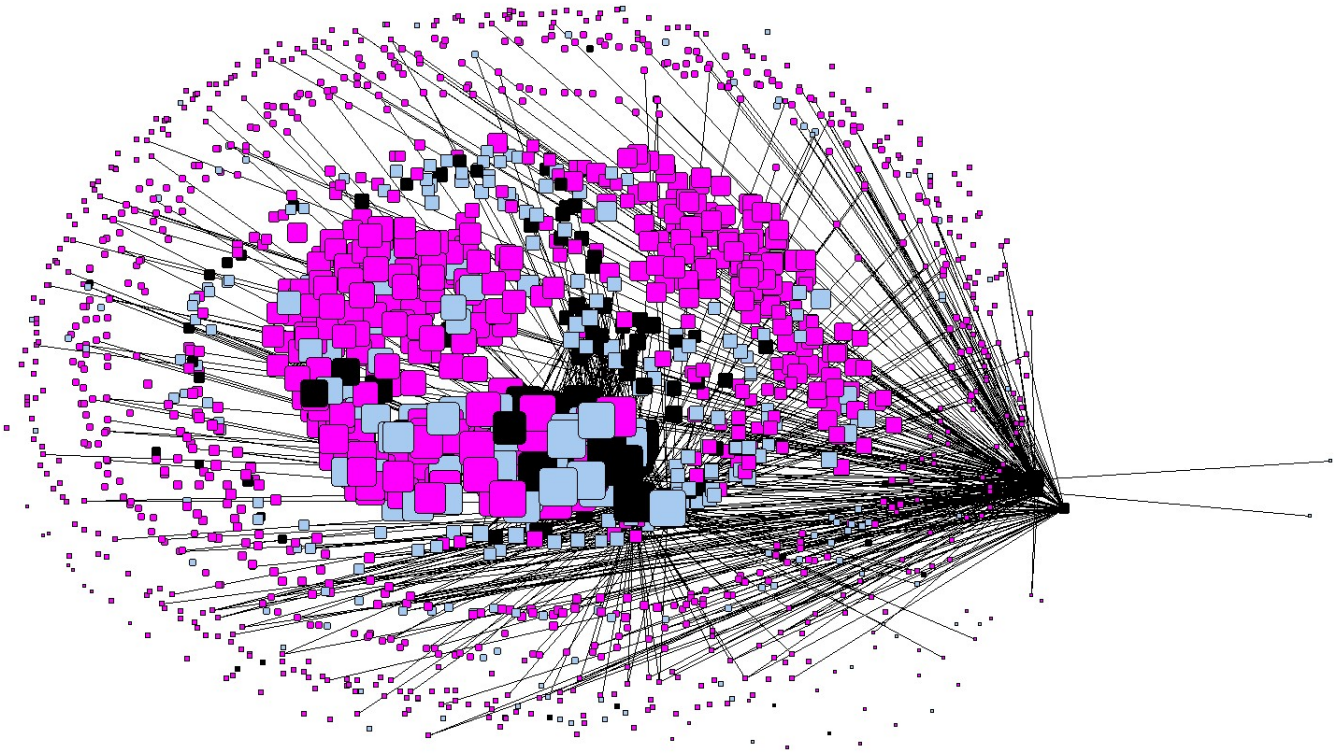


Figura 2.3: Rete monomodale attori/attori per tutti i domìni di esperimento - misura di degree centrality

La dimensione dei nodi è proporzionale alla loro degree centraliy, intesa qui come il numero di legami in ingresso/uscita. Si osservano con immediatezza due proprietà:

- il *core* vede la presenza di un rilevante insieme di nodi dotati di un significativo valore relativo di centralità, ad indicare l'ampiezza del sistema socio-produttivo, almeno con riferimento all'intero arco temporale assunto a riferimento;

- la degree centrality non presenta una correlazione statisticamente significativa con l'attributo di ruolo (come è qualificato il singolo attore nell'ambito Wikipedia, da user ad administrator). Il che è consistente con l'ipotesi della compresenza de due circuiti di regolazione ipotizzati.

Passando dalla misura di centralità di grado a quella di betweenness centrality (figura 2.4), il grafo presenta una rilevante semplificazione. La betweenness centrality esprime quanto ogni singolo nodo svolga la funzione di connettore fra nodi fra loro non direttamente legati. Essa è calcolata con riferimento ai percorsi geodetici, ovvero di minor lunghezza. Maggiore il valore di betweenness centrality di nodo, maggiore il numero di percorsi geodetici che lo attraversano. Come tale, la misura esprime in forma sintetica il rapporto fra ogni suo componente (nodo) e la struttura complessiva del network.

Come si può osservare anche visivamente, gli attori con ruoli formali (neri e grigi) hanno associati valori mediamente più elevati, pur a fronte di una rilevante presenza di Prosumer attivi prevalentemente nell'ambito del proprio dominio (verificato attraverso elaborazioni qui non dettagliate). Ciò è consistente con l'ipotesi di compresenza dei due circuiti di regolazione, valori elevati di betweenness centrality caratterizzando in particolare il core del network.

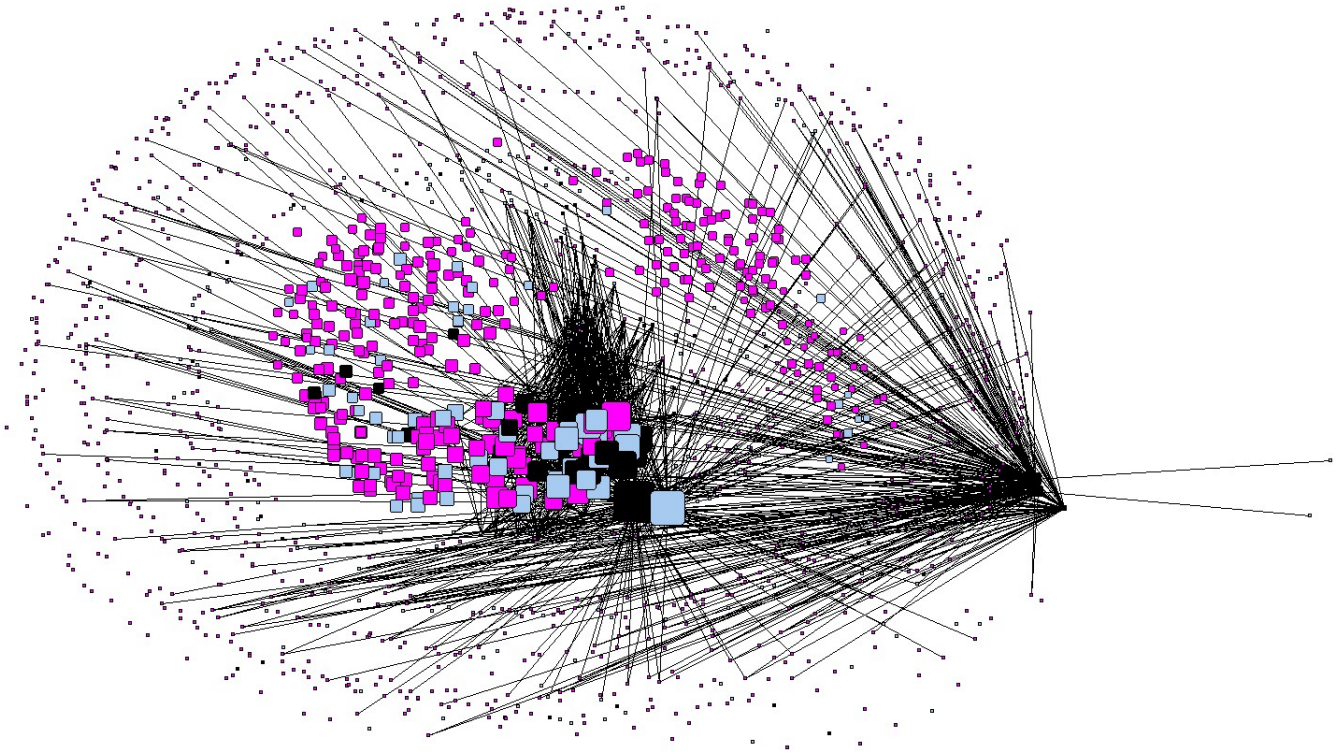


Figura 2.4: Rete monomodale attori/attori per tutti i domìni di esperimento - misura di betweenness centrality

2.3.2 Rete tra pagine

Passando alla rete monomodale "pagine/pagine" (figura 2.5), si osserva anche in questo caso l'attesa presenza di una macro-componente connessa che include tutte le pagine campionate. Essa è l'esito della "legatura" del sistema socio-produttivo sopra introdotto.

La successiva dicotomizzazione progressiva del network (ovvero la successiva eliminazione dei legami deboli, per valori di cut off crescenti) mostra una articolazione coerente (nei limiti della numerosità delle pagine in esame) con la topologia core-periphery. In figura 2.6 è proposto il network successivamente a rimozione dei legami con valore >= 7, mantenendo la visualizzazio-

ne dei nodi ora isolati. Se, come atteso, il subnet maggiormente resistente è quello relativo alle pagine del dominio Calcio, si osserva la permanenza nella macro-componente di nodi afferenti agli altri due domìni, in ragione delle connessioni "di forza media" che presentano con il primo. In termini di esperimento, la permanenza di legami fra domìni differenti è consistente con l'ipotesi che siano possibili meccanismi di correzione basati sul tracking, ad opera di agenti "centrali" con ruoli formali.

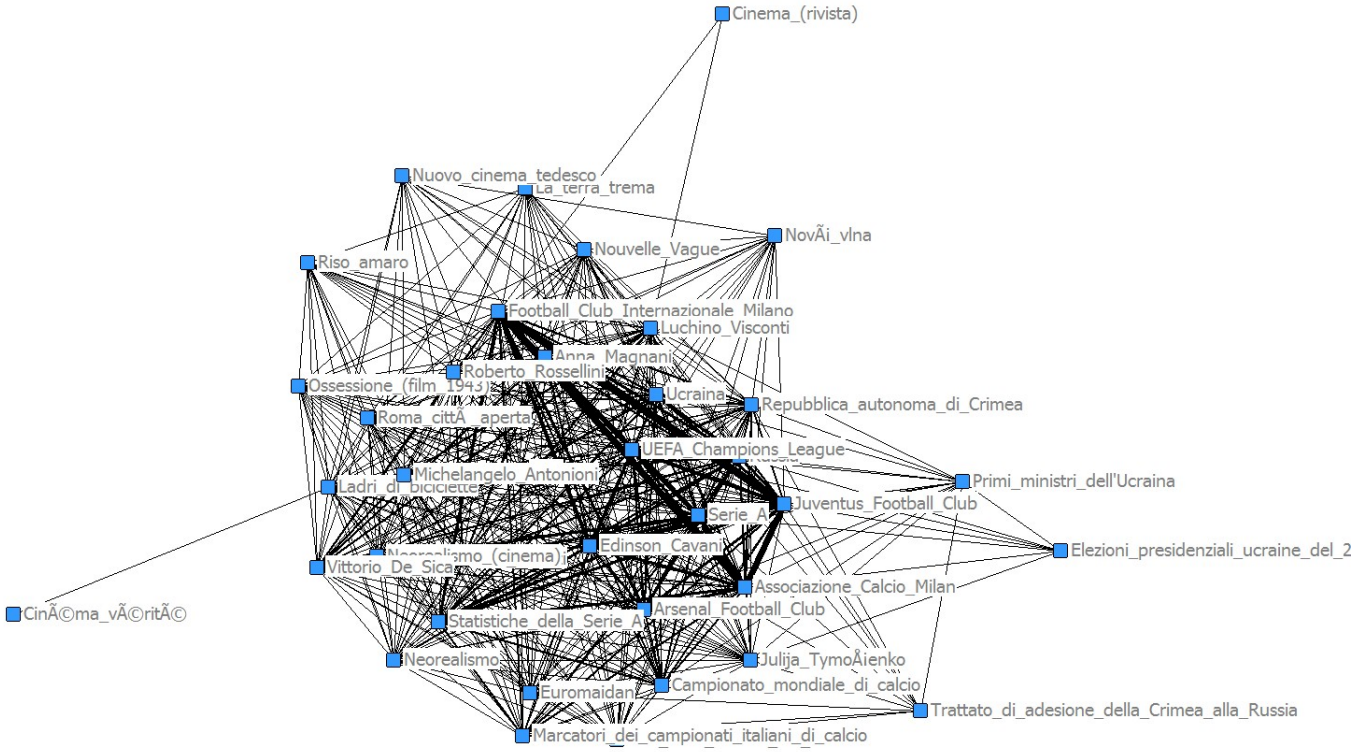


Figura 2.5: Rete monomodale pagine/pagine per tutti i domìni di esperimento - lo spessore dei legami è proporzionale al valore della relazione

Il differente valore dei legami (e, dunque, la più o meno estesa "sopravvivenza" delle pagine nel network, a mano a mano che viene condotta la dicotomizzazione) è inoltre positivamente utilizzabile come criterio guida per

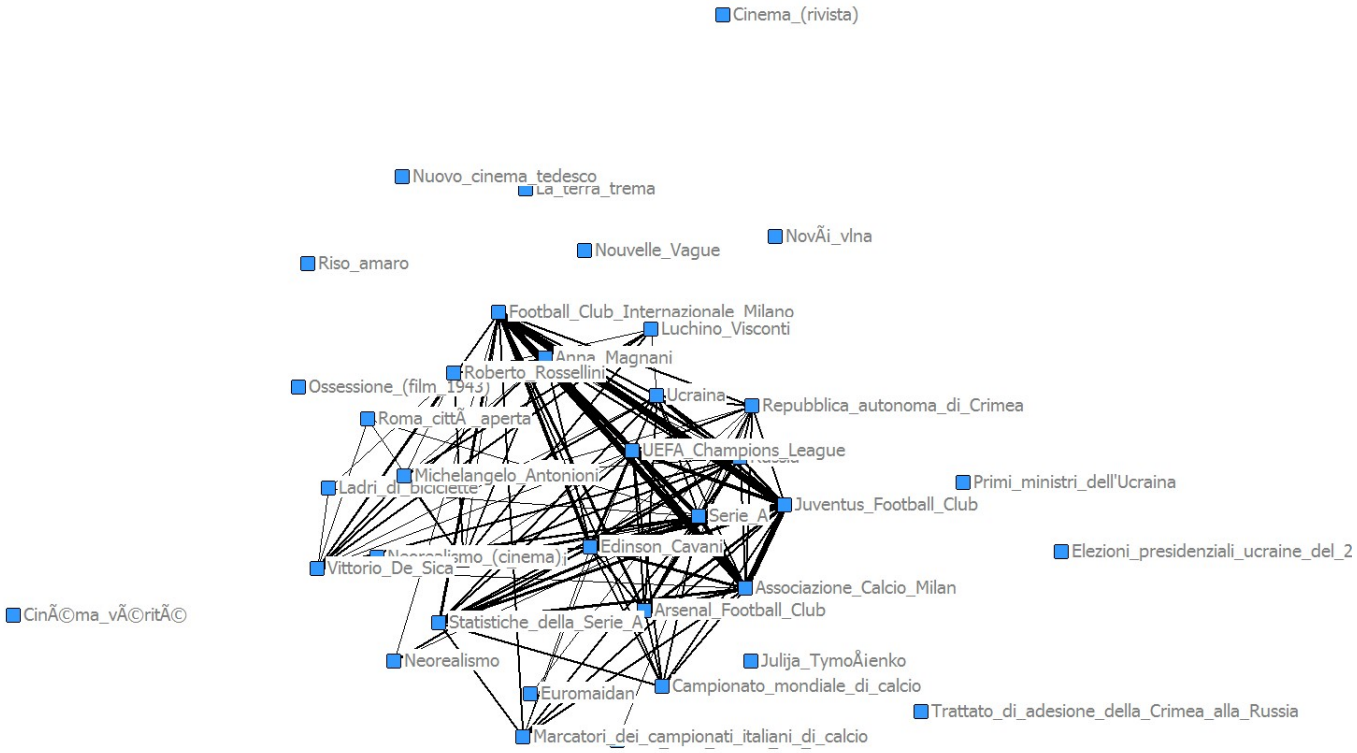
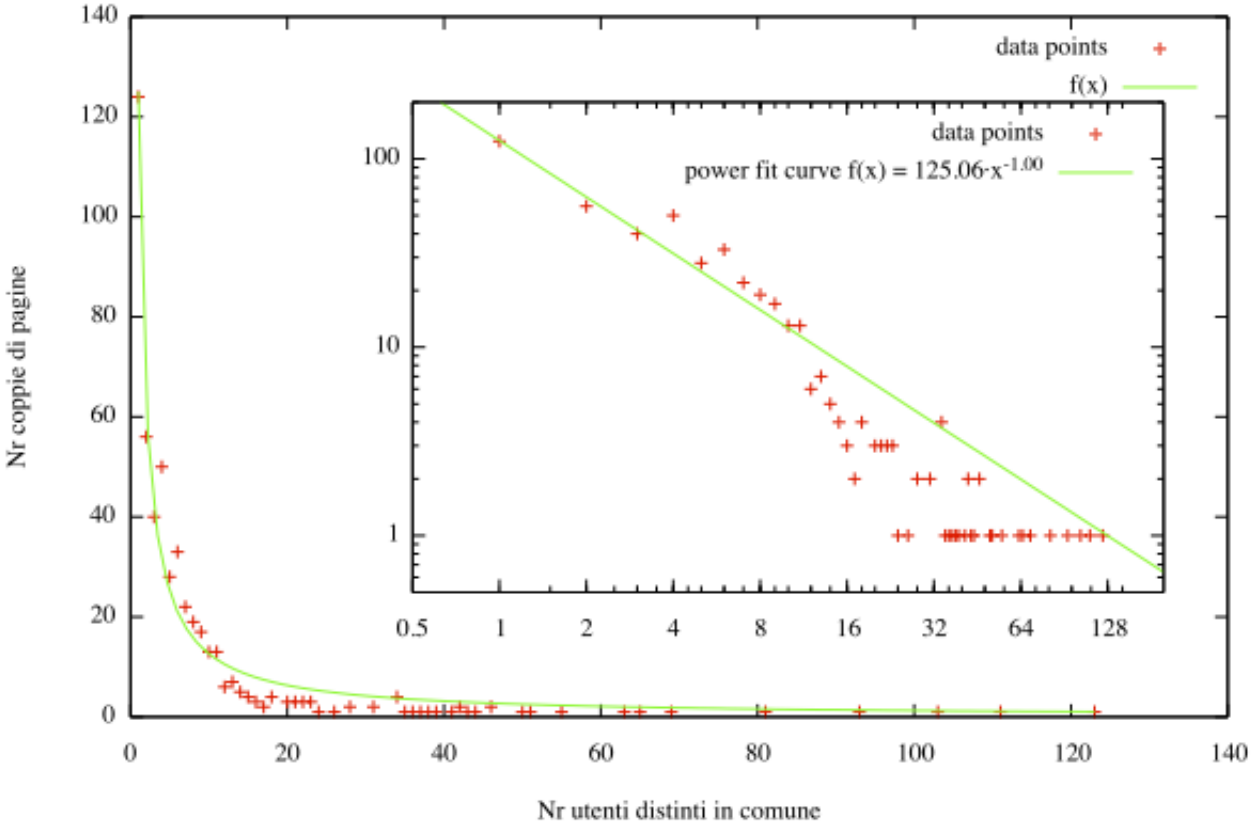


Figura 2.6: Rete monomodale pagine/pagine per tutti i domìni di esperimento. Dicotomizzazione GT7 - lo spessore dei legami è proporzionale al valore della relazione

segmentare le pagine stesse in categorie a differente probabilità attesa di correzione da parte sia della community, sia degli agenti "istituzionali". Un valore di legame forte indica un elevato numero di attori che operano su entrambi i nodi, accrescendo dunque la probabilità di rintracciamento sequenziale degli errori introdotti.

Nr occ	Coppie	Nr occ	Coppie	Nr occ	Coppie	Nr occ	Coppie
123	1	43	1	23	3	9	17
111	1	42	2	22	3	8	19
103	1	41	1	21	3	7	22
93	1	39	1	20	3	6	33
81	1	38	1	18	4	5	28
69	1	37	1	17	2	4	50
65	1	36	1	16	3	3	40
63	1	35	1	15	4	2	56
55	1	34	4	14	5	1	124
51	1	31	2	13	7		
50	1	28	2	12	6		
46	2	26	1	11	13		
44	1	24	1	10	13		

Tabella 2.7: Numero di utenti distinti in comune (Nr occ) per coppie di pagine (Coppie)



Occorrenze comuni	Nr. coppie di utenti	Occorrenze comuni	Nr. coppie di utenti
10	15	5	2459
9	37	4	5633
8	160	3	17010
7	361	2	87057
6	993	1	1152942

Tabella 2.8: Occorrenze comuni (pagine modificate da coppie di utenti) nel gruppo Calcio

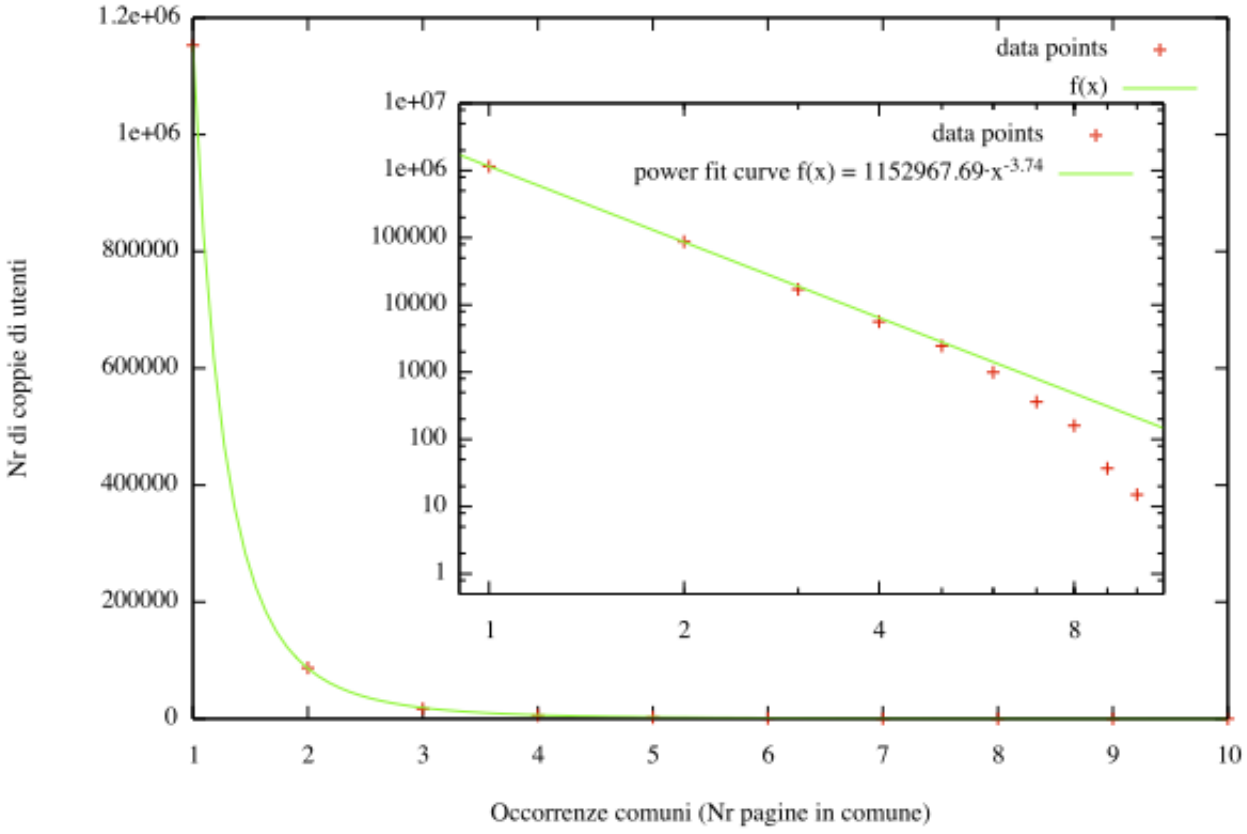


Tabella 2.9: Occorrenze comuni Calcio

Occorrenze comuni	Nr. coppie di utenti
5	9
4	157
3	735
2	5347
1	80681

Tabella 2.10: Occorrenze comuni (pagine modificate da coppie di utenti) nel gruppo Crisi ucraina

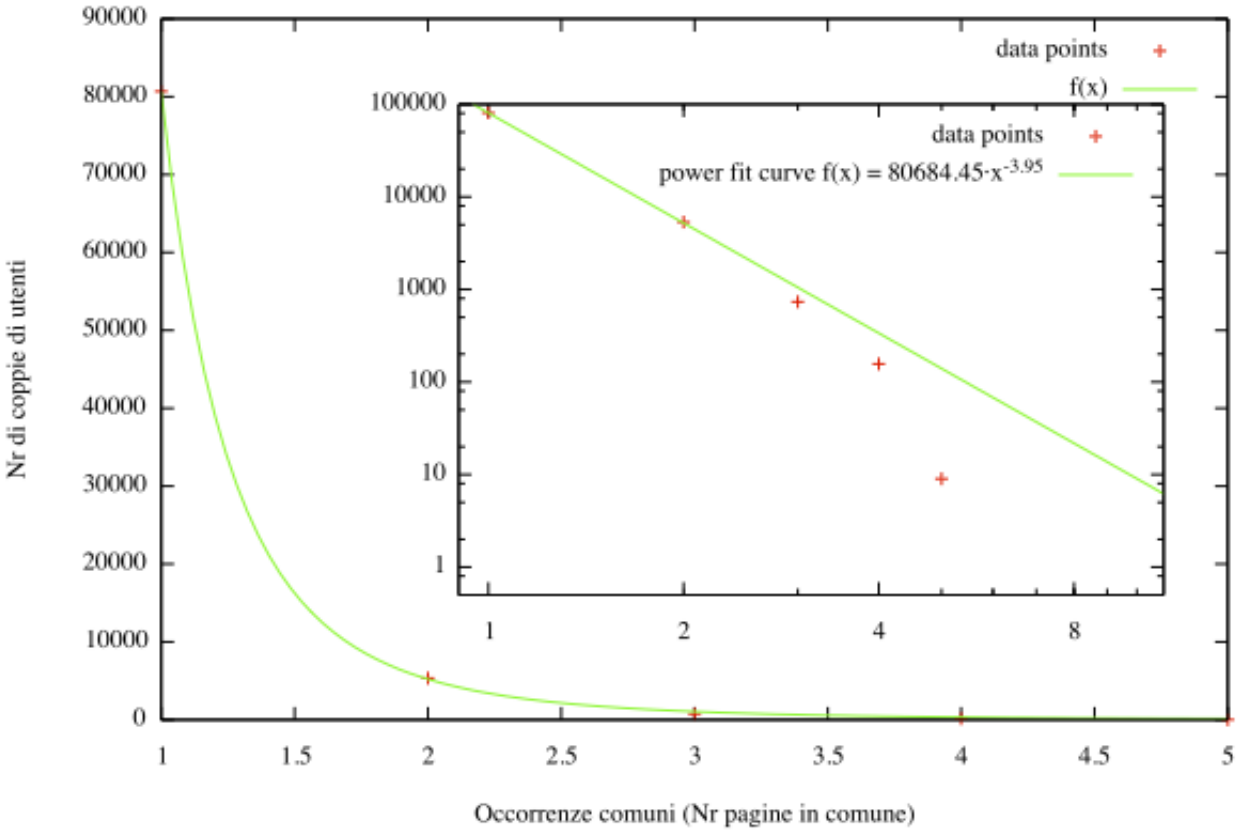


Tabella 2.11: Occorrenze Crisi ucraina

Occorrenze comuni	Nr. coppie di utenti	Occorrenze comuni	Nr. coppie di utenti
10	1	5	140
9	1	4	387
8	10	3	957
7	35	2	3517
6	60	1	38491

Tabella 2.12: Occorrenze comuni (pagine modificate da coppie di utenti) nel gruppo Neorealismo

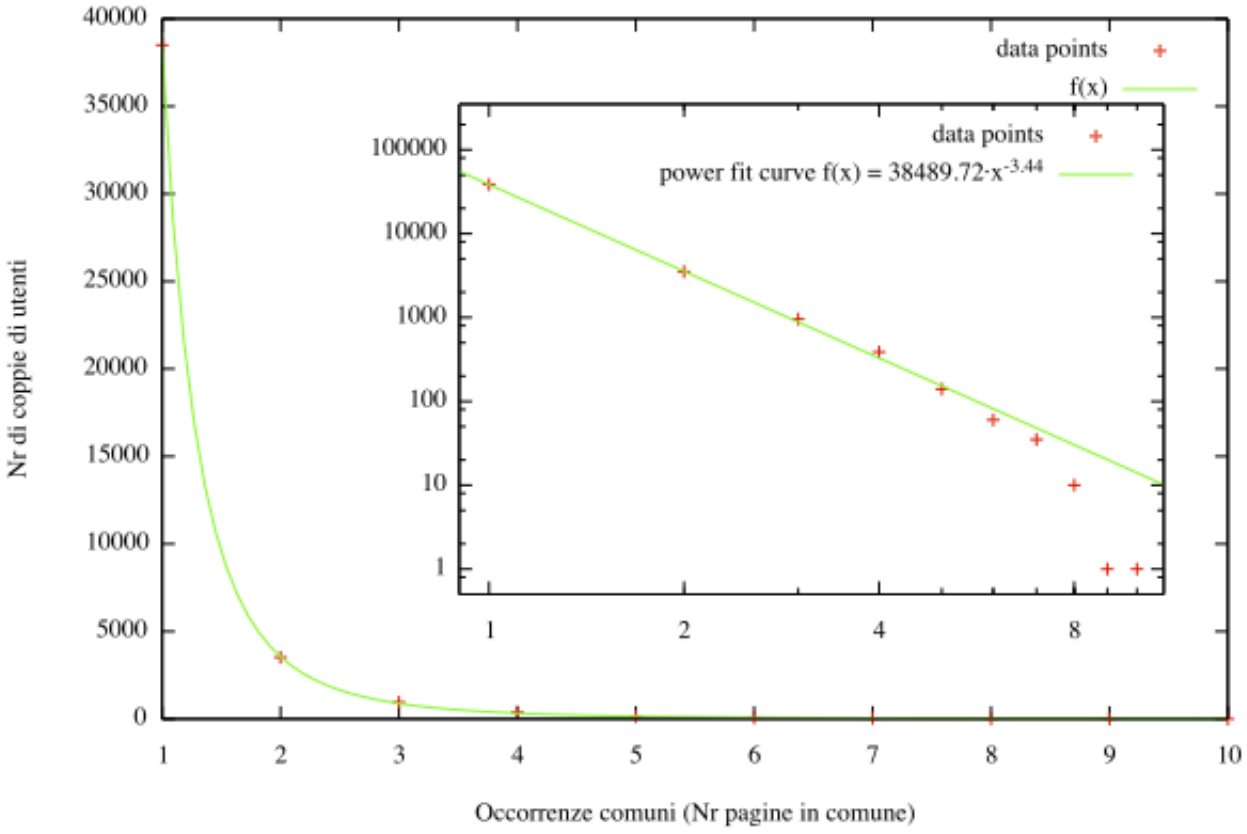


Tabella 2.13: Occorrenze comuni Neorealismo

Occorrenze comuni	Nr. coppie di utenti	Occorrenze comuni	Nr. coppie di utenti
18	1	8	395
16	4	7	761
15	6	6	1681
14	3	5	3564
13	9	4	7573
12	19	3	20615
11	37	2	96595
10	92	1	1243536
9	180		

Tabella 2.14: Occorrenze comuni (pagine modificate da coppie di utenti) sul totale di pagine

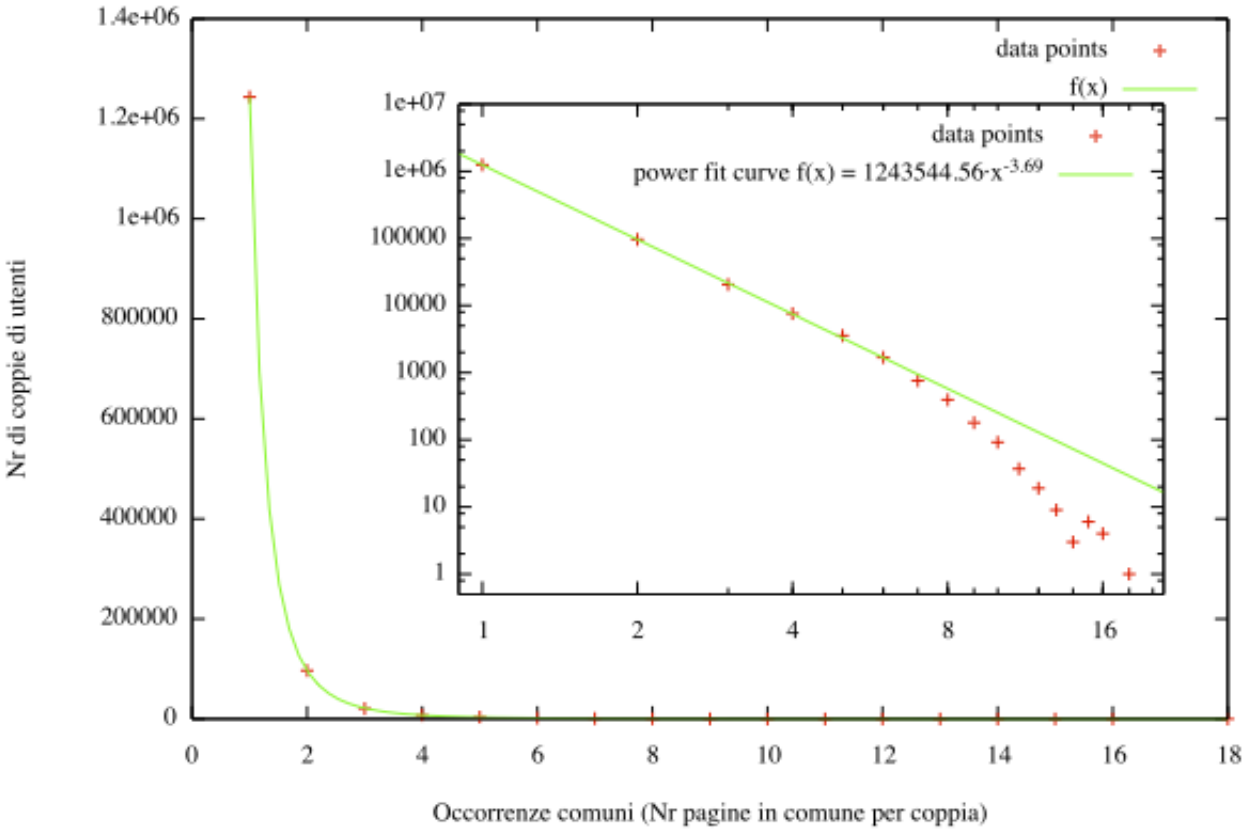


Tabella 2.15: Occorrenze totali

Nr. pagine modificate	Nr. utenti distinti	Nr. pagine modificate	Nr. utenti distinti
22	1	10	22
21	1	9	15
20	1	8	22
19	1	7	34
18	3	6	59
15	4	5	60
14	2	4	107
13	4	3	186
12	10	2	506
11	17	1	2343

Tabella 2.16: Nr di utenti distinti che hanno modificato un certo numero di pagine

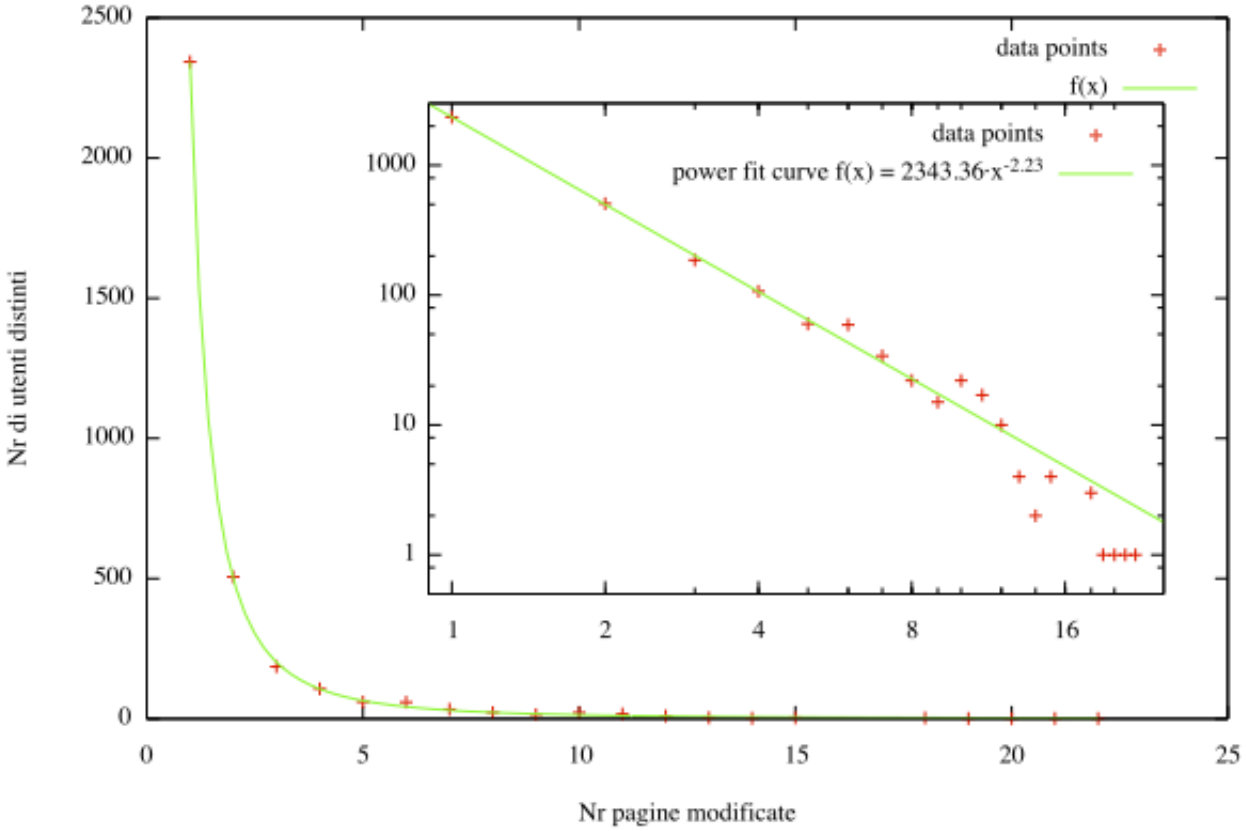


Tabella 2.17: Numero di pagine modificate da utenti distinti sul totale di pagine

Nr. pagine modificate	Nr. utenti distinti	Nr. pagine modificate	Nr. utenti distinti
10	6	5	42
9	6	4	83
8	16	3	145
7	19	2	421
6	35	1	1921

Tabella 2.18: Nr di utenti distinti che hanno modificato le pagine del gruppo Calcio

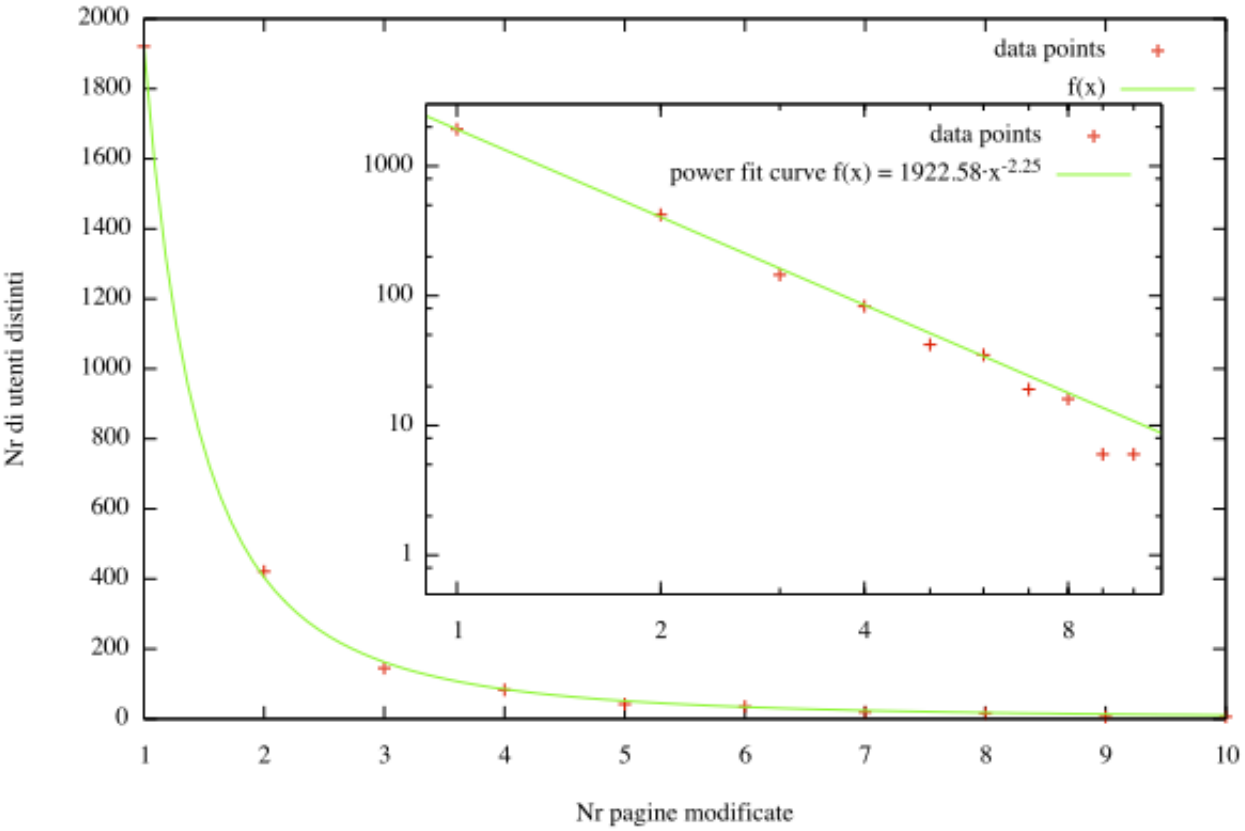


Tabella 2.19: Numero di pagine modificate da utenti distinti nel gruppo Calcio

Nr. pagine modificate	Nr. utenti distinti	Nr. pagine modificate	Nr. utenti distinti
11	3	5	11
10	3	4	30
9	4	3	36
8	4	2	90
7	9	1	405
6	8		

Tabella 2.20: Nr di utenti distinti che hanno modificato le pagine del gruppo Neorealismo

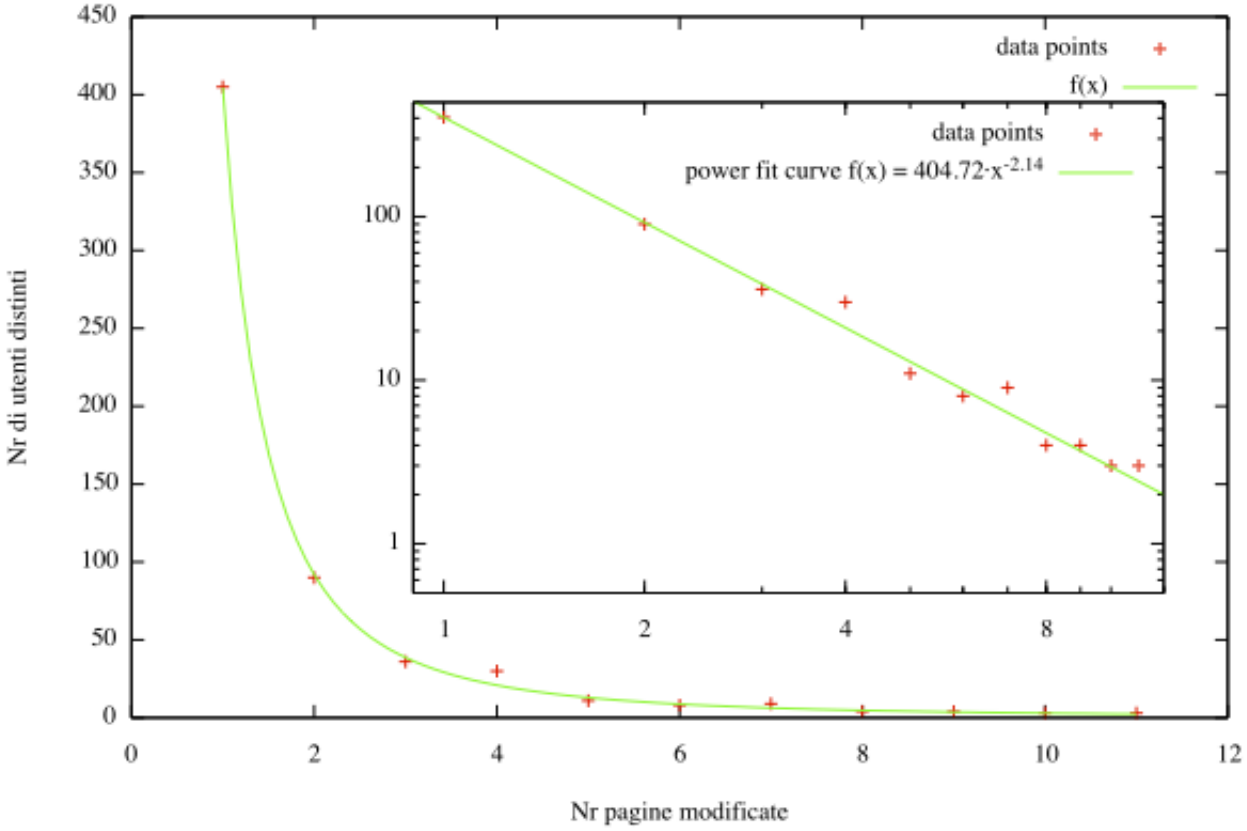


Tabella 2.21: Numero di pagine modificate da utenti distinti nel gruppo Neorealismo

Nr. pagine modificate	Nr. utenti distinti
6	4
5	9
4	15
3	40
2	93
1	455

Tabella 2.22: Nr di utenti distinti che hanno modificato le pagine del gruppo Crisi ucraina

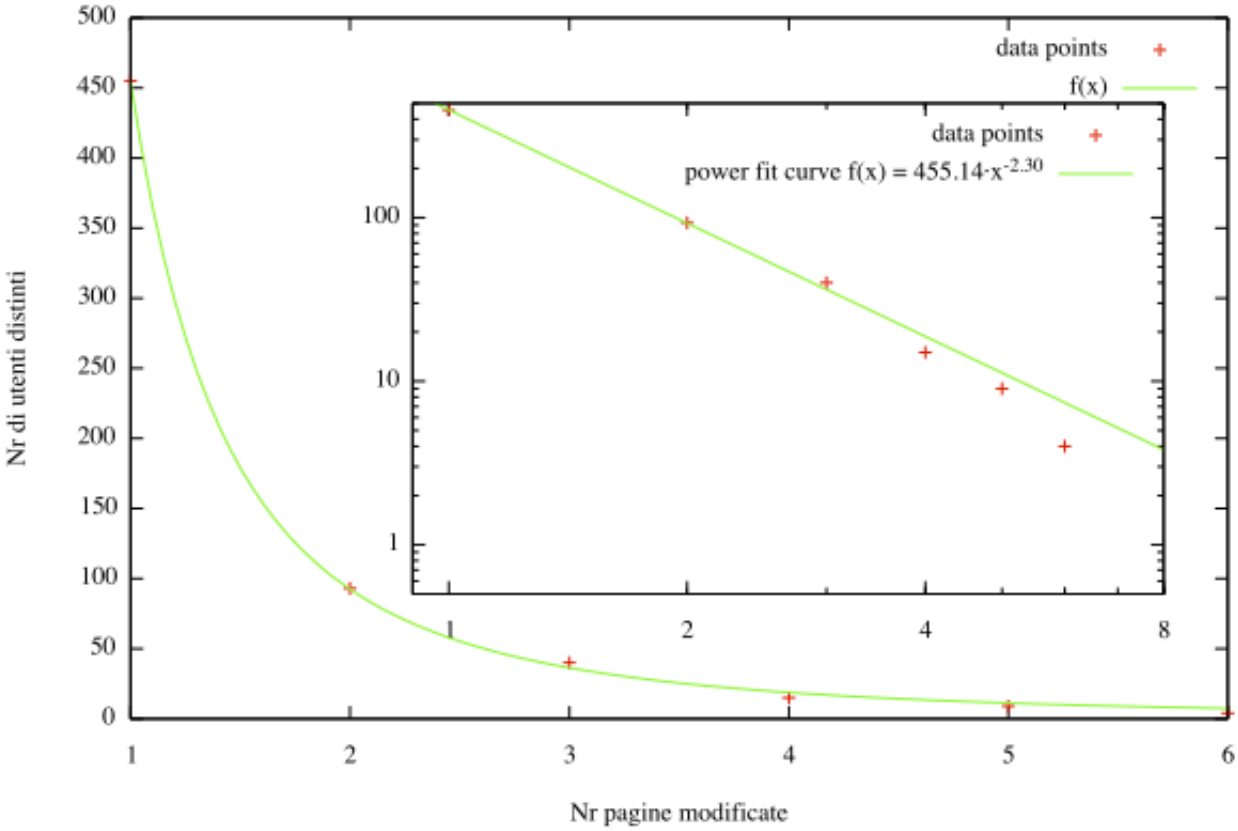


Tabella 2.23: Numero di pagine modificate da utenti distinti nel gruppo Crisi ucraina

Capitolo 3

Evidenze empiriche di autocorrezione

Per valutare le capacità autocorrettive del sistema, è stato scelto un approccio empirico che consiste nell'iniezione di errori all'interno delle pagine sotto specifici vincoli metodologici che sono descritti nelle prossime sezioni.

Vengono eseguiti due esperimenti tra loro indipendenti, volti a mostrare differenti aspetti di resilienza di Wikipedia. Entrambi gli esperimenti si appoggiano sul modello descritto nel capitolo precedente e rappresentato dalla figura 2.2. Come si vedrà in dettaglio nelle prossime sezioni, i due esperimenti differiscono principalmente per la tipologia di errori inseriti e per il metodo di iniezione utilizzato; oltre che l'analisi per la scelta delle pagine e delle tipologie di errori da iniettare, sarà quindi descritto anche lo studio relativo anche alla metodologia di inserimento degli errori.

3.1 Costruzione degli utenti

Per poter inserire gli errori, una settimana prima dell'avvio dell'esperimento sono stati appositamente creati degli utenti su Wikipedia, registrati con username e password, in modo tale che questi, passati 4 giorni dalla registrazione, acquisissero il ruolo *autoconfirmed*, diventando così utenti registrati

a tutti gli effetti e quindi abilitati alla modifica anche di pagine protette[1]. Per ciascuno di questi utenti, prima dell'inserimento degli errori, sono state effettuate delle modifiche produttive a Wikipedia, in pagine riferite ad argomenti vari, indipendenti dalle pagine selezionate per gli esperimenti. Questo approccio ha permesso di rendere gli utenti più "credibili" e non immediatamente identificabili come vandali. Tutte le iniezioni sono state effettuate manualmente per poter mantenere un controllo maggiore al momento dell'inserimento, utilizzando IP diversi per ciascun utente. Sono stati preparati 6 utenti, 3 per ciascun esperimento.

3.2 Esperimento 1

Lo scopo del primo esperimento è comprendere i meccanismi di autocorrezione del sistema verificando l'esistenza di circuiti di regolazione e analizzando come questi si comportano a seconda della categoria di pagine.

Per quanto riguarda il disegno dell'esperimento, si è scelto di utilizzare 3 utenti distinti, inserendo 3 errori di diversa entità per ciascuna pagina, su una selezione di 3 pagine per categoria, per un totale di 27 errori. Le variabili che influenzano questo disegno sperimentale sono l'approccio di inserimento degli errori, il grado di rilevanza (in termini di numero di legami) delle pagine oggetto dell'iniezione e la tipologia di errori inseriti.

L'ipotesi è che esista un sistema di autocorrezione all'interno di Wikipedia e che la correzione degli errori inseriti all'interno delle pagine appartenenti al nucleo della rete avvenga con maggior probabilità rispetto a quelle di pagine periferiche, in quanto più interconnesse. Da un punto di vista di schema di correzione, si ipotizza che gli errori siano corretti da parte dei Producer per quanto riguarda le pagine periferiche in quanto derivanti da una rilevazione dell'errore avvenuta tramite patrolling, poichè una correzione puntuale

[1]Per preservarne il contenuto, alcune pagine queste vengono segnalate da un amministratore come *protette* e sono modificabili solo da utenti registrati da più di 4 giorni. Generalmente questo accade per pagine spesso vittime di vandalismi

su pagine poco accedute da parte di un utente Prosumer risulta meno probabile, sebbene più precisa. Per quanto riguarda le pagine appartenenti al *core*, queste ci si aspetta che possano essere corrette sia da Producer che da Prosumer, con maggior probabilità rispetto alle pagine periferiche.

Come ipotesi secondaria si pensa che il meccanismo di autocorrezione potrà essere influenzato, oltre che dalla appartenenza o meno della pagina al core, anche dalla gravità degli errori inseriti.

3.2.1 Metodologia di iniezione degli errori

Il primo esperimento è caratterizzato dal fatto che ciascun utente inserisce gli errori su tutte le pagine della stessa categoria, come raffigurato in figura 3.1. Questo rapporto 1 a 1 tra utenti e categorie è sufficiente per mettere in evidenza la presenza di un certo tipo di circuito di regolazione e permette di ricavare se il comportamento di autocorrezione varia o si mantiene uguale nei 3 set sperimentali. La scelta di utilizzare 3 utenti differenti è per limitare l'association effect [10] a livello di categoria.

Al momento dell'inserimento degli errori, per ciascuna pagina si è cercato, ove possibile, di produrre, con lo stesso utente che utilizzato per iniettare gli errori, delle modifiche migliorative marginali sulla pagina (correzione della forma delle frasi, grammatica, ortografia, punteggiatura), allo scopo di rendere meno palese l'attività di vandalismo dell'utente.

Gli errori sono stati inseriti effettuando registrazioni indipendenti delle modifiche sulla pagina, ovvero "salvando" la pagina ad ogni modifica, di modo da evitare che tutti gli errori possano essere corretti tramite un unico revert.

3.2.2 Selezione delle pagine

Per selezionare le pagine oggetto degli esperimenti sono stati utilizzati i risultati delle analisi delle reti tra pagine descritte nel Capitolo 2, ricavando, per ciascuna categoria, 3 pagine di rilevanza crescente all'interno della

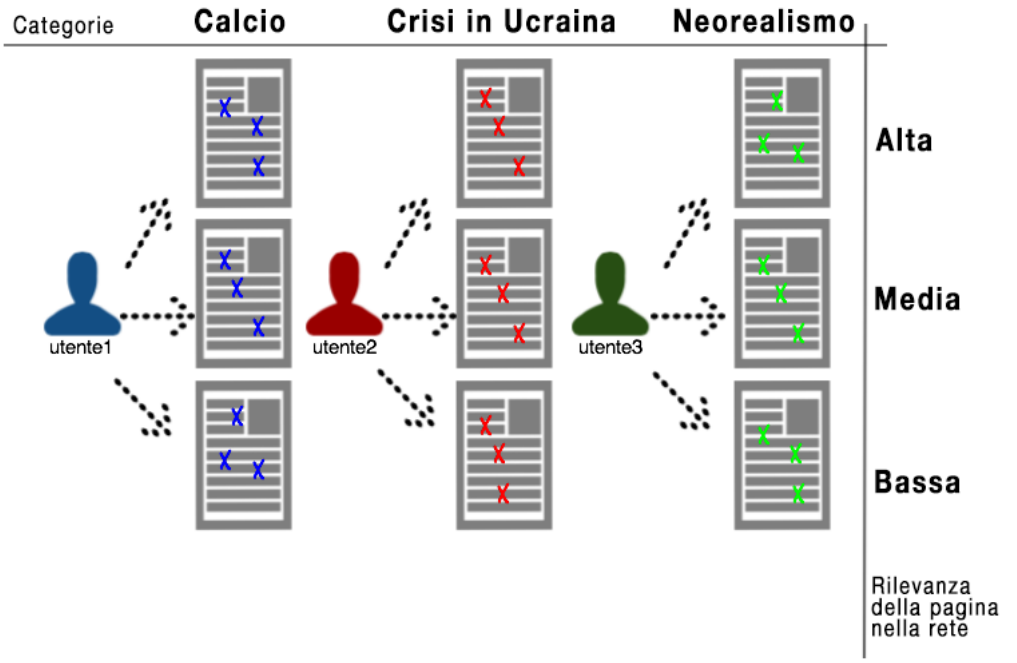


Figura 3.1: Metodologia di iniezione degli errori per l'esperimento 1, associazione 1:1 utenti-categorie

rete: una pagina statica (*rilevanza bassa*) con poche connessioni verso il *core*, una pagina intermedia e la pagina principale per la categoria (*rilevanza alta*) appartenente al core e quindi con molti legami. Il primo approccio di selezione delle pagine è quello dicotomico sul valore di n, che rappresenta il numero di legami comuni tra le pagine; aumentando n, ovvero il numero di legami minimo affinchè la rete mantenga una connessione tra i nodi, è stato possibile estrarre le pagine *minori* (le prime a perdere collegamenti con la rete all'aumentare di n) e quelle *intermedie* per le 3 categorie, anche se è importante evidenziare anche come il *core* della rete fosse quasi esclusivamente composto da pagine della categoria Calcio, per cui la dicotomizzazione ha portato risultati per le pagine del set Calcio solo al raggiungimento di alti valori di n mostrando la maggior stabilità della rete nel nucleo. A valori rispettivamente di $n > 12$ e $n > 14$, le intere categoria Neorealismo e Ucraina

perdono connessione con il resto della rete creando delle subnet.

Per quanto riguarda l'estrazione delle pagine principali di ogni categoria, la dicotomizzazione non è stata sufficiente per evidenziare queste pagine. Un filtro sul grado di betweenness non si è rivelato adatto alla selezione poichè estremamente influenzato dai nodi più periferici che possedevano esclusivamente un collegamento ad una pagina più interna e attraverso la quale potevano raggiungere l'intera rete, facendo così aumentare eccessivamente la rilevanza del nodo intermedio rispetto alla rete. Si è quindi scelto di applicare un filtro di centralità di grado sulla rete ad un certo livello di n, che ha portato all'evidenza delle pagine più rilevanti della rete per ciascuna categoria. Dall'analisi della struttura della rete si può definire un valore soglia di $n > 35$ per identificare le pagine appartenenti al *core*. Considerando questa soglia, si nota come siano appartenenti al *core*: la pagina principale e la pagina "intermedia" estratte per la categoria Calcio, la pagina principale per il set Crisi in Ucraina e nessuna delle pagine della categoria Neorealismo, in quanto per quest'ultima non è stato possibile ricavare una pagina con valore di legame sufficiente.

I valori di filtro corrispondono alla probabilità di una pagina di essere corretta, che è quindi correlata alla forma della rete, per cui per una pagina estratta ad $n > 6$, la probabilità di essere corretta sarà dimezzata rispetto ad una pagina estratta a $n > 12$. Si è scelto di escludere la pagine estratte a $n < 6$ in quanto troppo poco accedute e con probabilità di correzione troppo bassa. Per i nodi principali di categoria, inoltre, la probabilità di correzione aumenta in quanto le connessioni con altre pagine sono presenti anche inter-categoria.

Le pagine estratte sono riportate nella tabella 3.1.

3.2.3 Selezione degli errori

Per ciascuna pagina si è scelto di inserire 3 errori di tipologia differente: un errore grave, un errore medio ed uno di minor entità. Per la definizione degli errori si considerano due variabili ortogonali: la posizione dell'errore al-

Neorealismo	
Pagina	***n***
Riso amaro	6
Roma città aperta	12
Vittorio De Sica	Filtro di centralità di grado a $n = 12$
Crisi in Ucraina	
Pagina	***n***
Elezioni presidenziali ucraine del 2010	6
Kiev	12
Ucraina	Filtro di centralità di grado a $n = 12$
Calcio	
Pagina	***n***
Marcatori dei campionati italiani di calcio	16
Arsenal Football Club	44
Juventus Football Club	Filtro di centralità di grado a $n > 50$

Tabella 3.1: Pagine estratte - Esperimento 1.

l'interno nella pagina e la gravità dell'errore; il corretto equilibrio tra queste due variabili permette la definizione di errori di diversa tipologia. Un errore inserito in alto, tra le prime righe della pagina, lo si ipotizza più evidente per l'utente che accede in lettura alla pagina (questo approccio si applica tendenzialmente ai Prosumer, mentre per i Producer, che accedono tendenzialmente tramite patrolling basato su alert e sulla lista delle "Ultime modifiche", questa variabile sembra poco rilevante); ciò comporta una maggior probabilità per l'errore di essere corretto rispetto ad uno inserito nel testo di un paragrafo a fine pagina. Per giungere a questa idea si è considerato anche il fatto che gli utenti che consultano Wikipedia tendenzialmente lo facciano in quanto scarsamente esperti dell'argomento e spesso alla ricerca di informazioni di

base su questo, per cui risulta meno probabile la lettura dell'intera pagina rispetto invece alla raccolta di informazioni più generali ricavabili dalle parti principali della pagina.

Si è ipotizzato che un maggior numero di errori su ogni pagina avrebbe portato ad un'attenzione troppo elevata su quella pagina influenzando gli schemi di correzione. In un primo momento si era deciso di inserire anche errori *outlier*, estremamente evidenti e rilevanti, ma per evitare che le correzioni fossero troppo influenzate da questi errori si è scelto di evitarne l'inserimento.

In appendice B sono riportati gli errori selezionati per ciascuna pagina. Per motivi di leggibilità, in seguito si farà riferimento solo alla tipo di errore per ciascuna pagina.

3.3 Esperimento 2

Lo scopo del secondo esperimento è quello di verificare l'esistenza di schemi di tracking nella correzione degli errori. L'ipotesi sperimentale è quindi che esistano schemi di propagazione che permettono agli utenti Producer di correggere gli errori passando da una categoria all'altra seguendo le modifiche effettuate dall'utente (come rappresentato graficamente in Figura 3.2). In questo disegno sperimentale le variabili sono l'approccio di inserimento degli errori e il grado di importanza del pagine oggetto dell'iniezione.

Anche in questo caso si è scelto di utilizzare un approccio di iniezione basato su 3 utenti.

3.3.1 Procedimento per fasi

Per l'esperimento sono state sviluppate tre fasi successive, attivate dinamicamente durante l'esecuzione dell'esperimento a seconda della risposta del sistema.

Fase 1: prima iniezione. Vengono inseriti 2 errori di entità medio/bassa per ciascuna pagina, su 3 pagine per ciascuna categoria seguendo lo schema di selezione applicato per l'esperimento 1. Sotto-ipotesi: il sistema di autocorrezione si attiva indipendentemente dalla gravità degli errori inseriti.

Fase 2: seconda iniezione. L'approccio della seconda fase consiste ancora nell'iniezione di 1 errore di entità medio/bassa ma su pagine per le quali vi sia stata evidenza dell'esistenza di un sistema di autocorrezione.

Fase 3: terza iniezione. Aggiunta la variabile gravità dell'errore inserendo 1 errore di grande entità sulle pagine che si ipotizza vengano corrette con maggiore probabilità. Sotto-ipotesi: il sistema di autocorrezione si attiva solo per errori di grande entità.

Sebbene suddiviso in 3 fasi, lo scopo dell'esperimento resta comune: verificare l'esistenza di schemi di tracking e propagazione nella correzione degli errori. Nella fase 1 quindi, la potenziale esistenza di uno schema di tracking sarebbe l'evidenza che la gravità dell'errore non influisce sui sistemi di autocorrezione di Wikipedia. Al fallimento della fase 1, ovvero la mancanza di evidenza di schemi di propagazione, si procede alla seconda fase, volta a far attivare gli schemi di tracking focalizzando l'inserimento su pagine per le quali è già stata provato il funzionamento di un sistema autocorrettivo; anche in questo caso la gravità dell'errore non sarebbe rilevante. Al fallimento anche della seconda fase, si procede con la fase 3, che inserisce la variabile gravità dell'errore come elemento chiave per l'attivazione dei sistemi di autocorrezione.

Allo scatenarsi delle correzioni si potranno quindi evidenziare i processi di tracking, studiandone il comportamento in maniera globale su tutto l'esperimento.

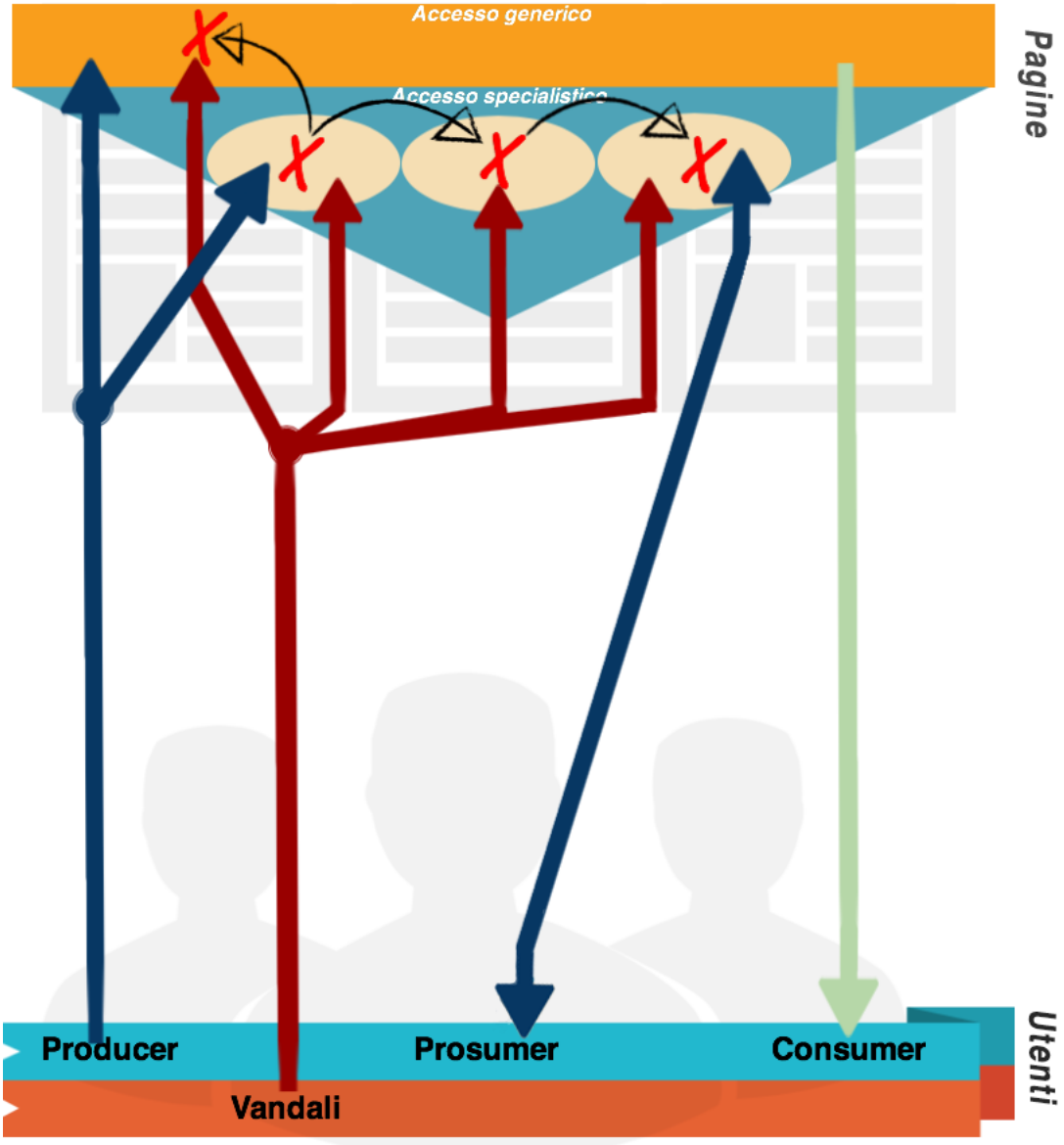


Figura 3.2: Esperimento 2 applicato al modello teorico di autocorrezione. Le linee curve nere rappresentano lo schema di tracking degli errori ipotizzato per i Producer

3.3.2 Metodologia di iniezione degli errori

In questo esperimento la variabile dipendente è la modalità di tracking dei supervisori, per cui l'iniezione degli errori avviene in maniera *incrociata*: ciascun utente inserisce errori su una pagina di ciascuna categoria, accomunate dal grado di rilevanza della pagina all'interno della rete (una pagina "importante" con molti legami di rete, una pagina "intermedia", una pagina "marginale" con pochi legami, sempre in riferimento alla categoria di appartenenza) come raffigurato in figura 3.3. Anche nelle le fasi 2 e 3 si è mantenuto questo schema di inserimento: gli errori sono stati iniettati con associazione 1:1 tra utenti e grado di rilevanza della pagina, ma non è stato più necessario incrociare le categorie poichè per scelta sperimentale in queste due fasi si è scelto di limitare l'inserimento alla categoria Calcio.

Anche per questo esperimento si è cercato di apportare modifiche migliorative marginali alle pagine al momento dell'inserimento degli errori e di inserirli "salvando" la pagina ad ogni modifica.

3.3.3 Selezione delle pagine

Fase 1

La selezione delle pagine in questo caso è avvenuta per somiglianza in termini statistici rispetto alle pagine precedentemente selezionate per il primo esperimento, così da presentare una base di pagine confrontabile per i due esperimenti. Il valore statistico considerato per il confronto è il numero di accessi in modifica per le pagine dalla data di creazione alla data di estrazione dei valori (Maggio 2014), come riportati nelle tabelle 2.1, 2.3, 2.2; inoltre si è tenuto conto anche del numero di accessi in modifica ricevuti da quelle pagine nell'ultimo anno e del numero di utenti che questa pagina ha in comune con pagine di altre categorie.

Per quanto riguarda la pagina di minor rilevanza della categoria Calcio, poichè il numero di revision non è stato sufficiente a dimostrare la somiglianza della pagina con la corrispondente dell'esperimento 1, è stato confrontato

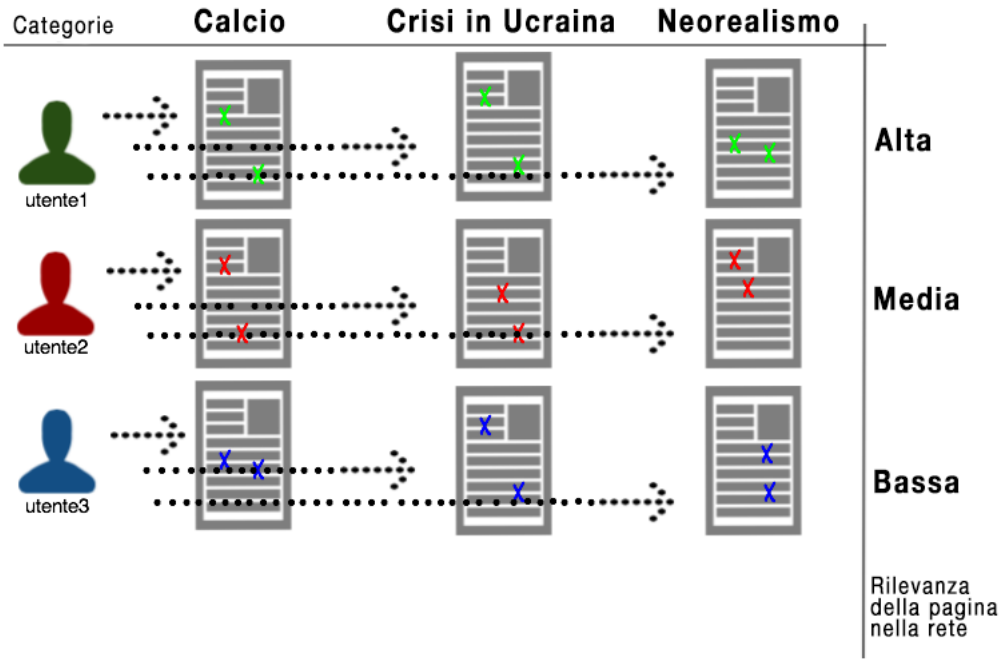


Figura 3.3: Metodologia di iniezione degli errori per l'esperimento 2, inserimento incrociato: una pagina di ciascuna categoria per ogni utente

il dato con l'analisi dicotomica descritta per il primo esperimento che ha confermato che la pagina selezionata è effettivamente tra le prime della categoria Calcio a perdere connessione con la rete a $n > 16$.

Le pagine estratte sono riportate nella tabella 3.2.

Fase 2

Le pagine selezionate per la seconda fase dell'esperimento sono le pagine per le quali è già stata verificata l'esistenza di un sistema autocorrettivo. Poichè i due esperimenti sono stati eseguiti in parallelo, la seconda fase dell'esperimento 2 ha potuto sfruttare i risultati parziali dell'esperimento 1, ovvero le evidenze di correzione sulle pagine della categoria Calcio dell'esperimento 1 (tabella 3.3) . Queste pagine, come mostrato nella tabella riassuntiva dei risultati dell'esperimento 1 3.4, sono state corrette entro la prima ora e si

Neorealismo
La terra trema
Neorealismo
Anna Magnani
Crisi in Ucraina
Primi ministri dell'Ucraina
Repubblica autonoma di Crimea
Russia
Calcio
Campionato mondiale di calcio
UEFA Champions League
Football Club Internazionale Milano

Tabella 3.2: Pagine estratte - Esperimento 2, fase 1

possono ritenere un risultato congelato a livello dell'esperimento 1, per cui è possibile riutilizzarle senza influenzarne gli sviluppi successivi. Questa fase è attivata 2 ordini di grandezza temporali dopo la rilevazione della correzione sulle pagine interessate dell'esperimento 1 (circa 100 ore dopo la prima correzione).

Calcio
Marcatori dei campionati italiani di calcio
Arsenal Football Club
Juventus Football Club

Tabella 3.3: Pagine estratte per la categoria Calcio - Esperimento 2, fase 2

Fase 3

Le pagine per la terza fase dell'esperimento 2 sono le stesse della fase 1, limitatamente a quelle della categoria Calcio poichè, appartenendo al *core* della rete, sono le pagine la cui probabilità di essere corrette è maggiore.

3.3.4 Selezione degli errori

Anche in questo caso gli errori sono stati generati tenendo conto delle due variabili: posizione dell'errore e gravità dell'errore. In questo modello sperimentale si è scelto di utilizzare 2 errori di entità media per la prima fase, 1 errore di entità media per la seconda fase e infine 1 errore grave per la terza fase.

In appendice B sono riportati gli errori selezionati per ciascuna pagina. Per motivi di leggibilità, in seguito si farà riferimento solo alla tipo dell'errore per ciascuna pagina.

3.4 Risultati

Per ciascun errore di entrambi gli esperimenti è stato misurato l'intervallo di tempo che trascorre dal momento del suo inserimento sino alla sua correzione. La finestra temporale di analisi è stata di 30 giorni per l'esperimento 1 e di 23 giorni per la prima fase dell'esperimento 2, 18 per la seconda e 11 per la terza. Dalle statistiche sul tempo medio che intercorre tra una modifica e l'altra per ciascuna pagina utilizzata nell'esperimento, si può affermare che i risultati non siano stati influenzati dall'utilizzo di una finestra temporale troppo ristretta per quanto riguarda la presenza degli errori nelle pagine[1].

Per ciascun errore sono stati registrati i dati associati all'intervento di modifica: tipologia dell'utente che l'ha effettuato e metodo di correzione dell'utente (tramite revert o tramite una correzione puntuale dell'errore); sono stati anche riportati i valori del numero di visualizzazioni e numero di modifiche avvenute sulla pagina da parte di altri utenti in presenza dell'errore.

La legenda per la lettura delle tabella riassuntive delle correzioni per ciascuna pagina è la seguente:

[1]Valore *Mean time between edits* recuperato per ciascuna pagina da `http://vs.aka-online.de/cgi-bin/wppagehiststat.pl`

LE: Longevità dell'errore (in ore)

nV: Numero visualizzazioni della pagina dal momento dell'inserimento dell'errore

nM: Numero di modifiche effettuate sulle pagine dal momento dell'inserimento dell'errore

TC: Tipo correzione:

- **S-D:** Modifica eseguita da parte di un supervisore direttamente
- **S-T:** Modifica eseguita da parte di un supervisore seguendo un tracking da altre pagine
- **U-D:** Modifica eseguita da parte di un utente normale direttamente
- **U-T:** Modifica eseguita da parte di un utente normale seguendo un tracking da altre pagine

MC: Modalità di correzione:

- **R:** Revert
- **P:** Correzione puntuale

3.4.1 Esperimento 1

I risultati dell'esperimento 1 sono riassunti nella tabella 3.4.

Per alcuni errori, il valore di longevità è stato riportato come corrispondente all'intera finestra temporale dell'esperimento, ovvero 30 giorni, poichè per questi errori non è stata registrata alcuna correzione. Il 33% degli errori inseriti è stato corretto nella finestra temporale; si evidenzia però come questi corrispondano unicamente agli errori della categoria Calcio, le cui pagine sono state corrette nel giro di 1 ora da parte dello stesso utente supervisore.

In particolare:

- Juventus Football Club: corretta 24 minuti dopo l'inserimento dell'ultima modifica. Commento sulla revert: *info false/senza fonti*

- Arsenal Football Club: corretta 40 minuti dopo l'inserimento dell'ultima modifica. Commento sulla revert:*senza riscontri*

- Marcatori dei campionati italiani di calcio: corretta 57 minuti dopo l'inserimento dell'ultima modifica. Commento sulla revert: -*vandalismi*

Durante la correzione sono state eliminate anche le modifiche migliorative applicate alla pagina in quanto il supervisore ha effettuato un *revert* di tutte le modifiche eseguite da quell'utente sulla pagina.

Questo risultato è consistente con l'ipotesi principale riguardante l'esistenza di un sistema di autocorrezione dall'interno di Wikipedia e conferma la grande stabilità della rete nel nucleo. Per quanto riguarda invece gli errori inseriti nelle pagine appartenenti alle categorie Crisi in Ucraina e Neorealismo, il comportamento di correzione si è mostrato differente e in totale contrasto con il precedente: dopo 30 giorni dall'inserimento degli errori, questi non sono stati corretti, confermando le scarse capacità di autocorrezione man mano che ci si allontana dal nucleo della rete, in accordo con l'ipotesi di partenza.

Per quanto riguarda le ipotesi sullo schema di correzione, non è stato però possibile verificarle in modo esaustivo in quanto tutti gli errori sono stati corretti in blocco da un utente Producer. Si ipotizza che l'errore di maggior gravità inserito sulla pagina Juventus (la prima ad essere stata corretta e pagina con più alti valori di legame nella rete) abbia fatto nascere sospetti da parte del supervisore nei confronti dell'utente che ha inserito le modifiche, per cui è stato semplice per il Producer controllare le ultime modifiche effettuate dallo stesso vandalo sulle altre pagine ed effettuarne il revert in blocco. In questi termini si spiega anche il fatto che la pagina più distante dal nucleo della categoria Calcio (*Marcatori dei campionati italiani di calcio*) sia stata corretta mentre non è stato così per pagine delle altre categorie sebbene possedessero maggiori legami di rete (*Ucraina*, *Vittorio De Sica*). Si evidenzia quindi verosimilmente un sistema di tracking intra-categoria sull'utente malevolo.

3.4.2 Esperimento 2

I risultati dell'esperimento 2 sono riassunti nelle tabelle 3.5,3.6,3.7, rispettivamente per le fasi 1, 2 e 3.

Per alcuni errori, il valore di longevità è stato riportato come corrispondente all'intera finestra temporale dell'esperimento o della specifica fase, poichè per questi errori non è stata registrata alcuna correzione.

Il 21% degli errori inseriti è stato corretto nella finestra temporale ma nessuna delle 3 fasi dell'esperimento ha portato evidenze per quanto riguarda gli schemi di tracking, in quanto le uniche correzioni effettuate sono state eseguite in maniera puntuale, sia da parte di Prosumer che da parte di Producer e senza che si scatenassero meccanismi di propagazione delle correzioni; non vi sono quindi stati i presupposti per verificare l'ipotesi sperimentale.

É stato però possibile utilizzare i risultati di questo esperimento in correlazione con quelli del primo per estendere le valutazioni sui comportamenti autocorrettivi del sistema, come sarà esposto esaurientemente nella prossima sezione.

3.4.3 Combinazione e confronto dei risultati sperimentali

Per quanto riguarda la valutazione delle proprietà autocorrettive di Wikipedia, dalla combinazione dei risultati dei due esperimenti, risulta che soltanto 27% degli errori inseriti sono stati corretti nella finestra temporale (figura 3.4), il che indica una generale debolezza del sistema ai vandalismi, risultato che merita un'analisi dettagliata.

Come descritto precedentemente, dall'esperimento 1 si nota come gli errori della categoria Calcio vengano corretti nell'immediato, mostrando l'esistenza di un legame tra centralità della pagina nella rete e correzioni degli errori: gli errori appartenenti al *core* vengono corretti con maggiori probabilità. Dal confronto dei risultati dei due esperimenti si nota però che gli errori inseriti nella categoria Calcio durante l'esperimento 2 non sono stati

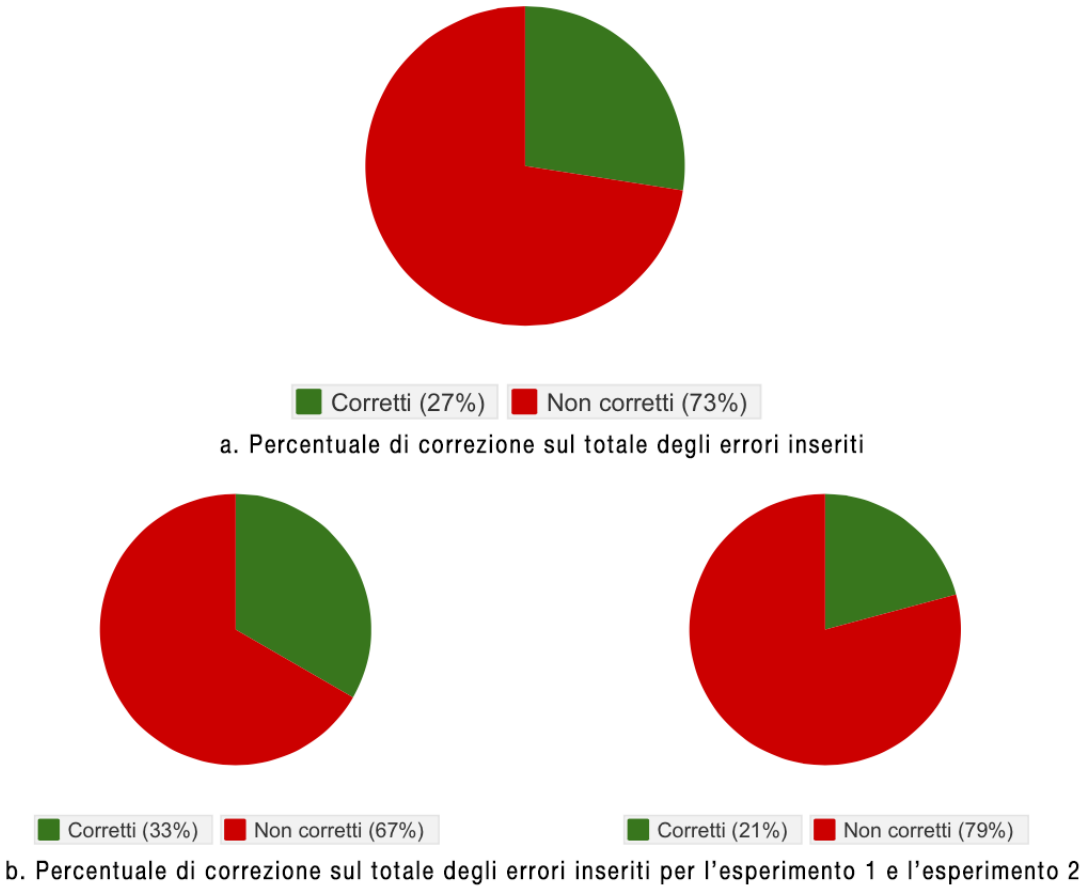


Figura 3.4: Distribuzione delle correzioni sul totale degli errori inseriti (a) e per esperimento (b)

corretti con le stesse tempistiche evidenziate durante l'esperimento 1 per la stessa categoria, nonostante le pagine selezionate per il secondo esperimento si possano definire equivalenti a quelle del primo (durante la prima e terza fase) o addirittura siano le stesse (durante la seconda fase).

Questo diverso comportamento si spiega pensando all'esistenza di una interdipendenza tra correzione degli errori ed entità di questi, che nelle prime due fasi dell'esperimento 2 sono stati tutti di entità medio/piccola. Appare quindi che per errori gravi la probabilità di correzione aumenta e portando alla correzione degli altri errori di minor entità sulla stessa pagina inseriti

contestualmente da quell'utente, mentre errori medio/piccoli non vengono corretti in maniera diretta anche nelle pagine del *core* (caso degli errori inseriti sulle pagine *Football Club Internazionale Milano*, *UEFA Champions League* e *Campionato mondiale di calcio* durante la prima fase dell'esperimento 2 e *Juventus Football Clu*b, *Arsenal Football Club* e *Marcatori dei campionati italiani di calcio* per la seconda fase).

La distribuzione delle correzioni evidenzia i seguenti risultati, rappresentati in figura 3.5 :

- da un punto di vista di tipologia di pagine, il 37% degli errori in pagine appartenenti al core sono stati corretti, contro il 22% di correzioni sulle pagine periphery.
- da un punto di vista di entità degli errori, il 42% degli errori gravi sono stati corretti, contro il 23% degli errori di entità media e piccola.

Osservando le correzioni combinando i criteri *entità dell'errore* e *rilevanza della pagina* emerge che sono stati corretti (figura 3.6) :

- il 60% degli errori gravi inseriti in pagine *core*
- il 29% di errori medio/piccoli inseriti su pagine *core*
- il 29% degli errori gravi inseriti nelle pagine *periphery*
- il 20% degli errori medio/piccoli inseriti su pagine *periphery*

Da ciò si nota dal punto di vista empirico una maggiore performance correttiva per errori gravi su pagine appartenenti al core, senza però che da ciò discendano meccanismi di correzione mediante tracking. Non è dunque possibile affermare che l'iniezione di un errore grave in una pagina rilevante porti ad un miglioramento generale della "tenuta" del sistema, nel caso di un attacco vandalico distribuito su pagine periphery, con diverse entità di danno.

Per analizzare i tempi di persistenza degli errori e quindi i tempi di correzione, è stato necessario focalizzarsi solo sugli errori che sono stati corretti

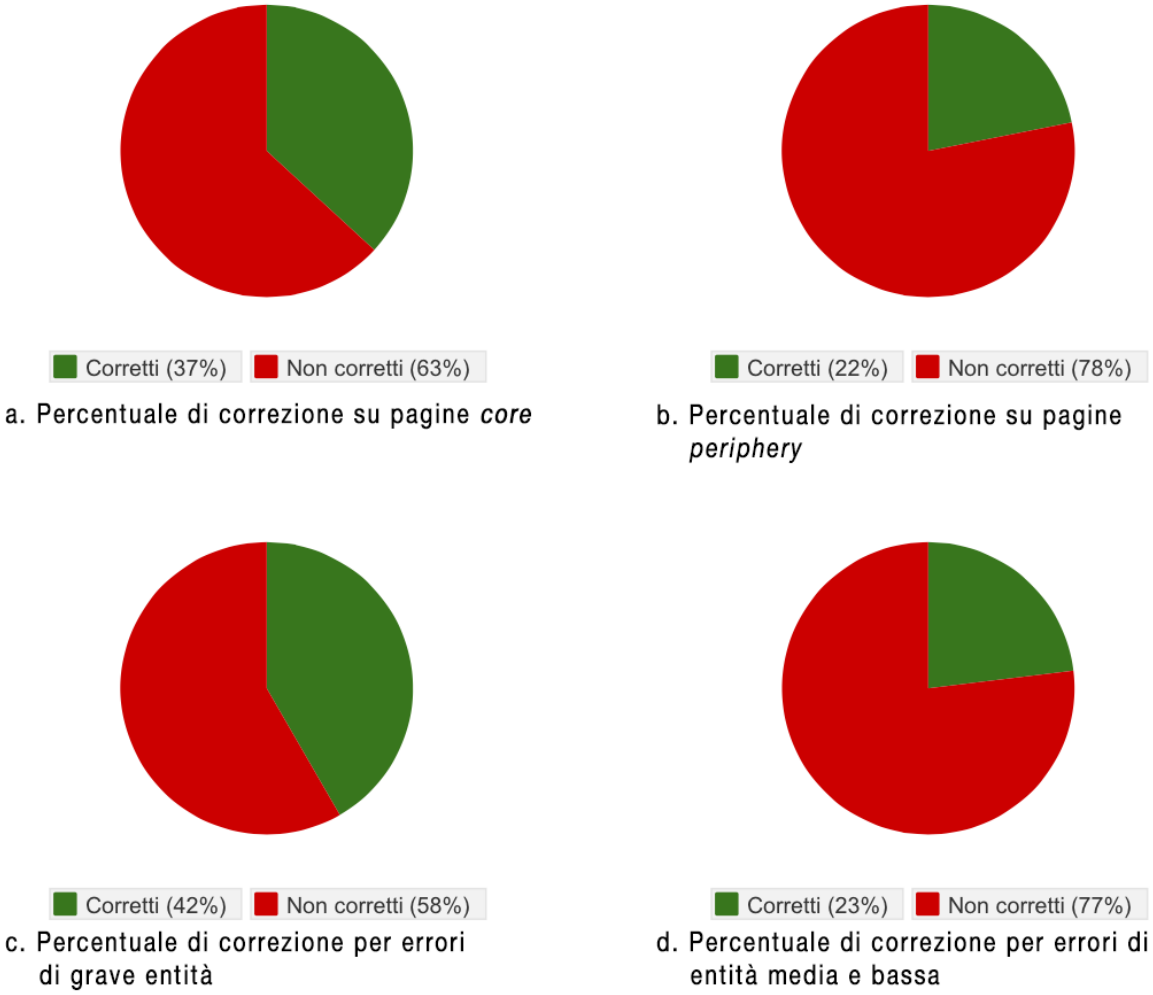


Figura 3.5: Distribuzione delle correzioni suddivise per tipologia di pagina (a,b) e per entità di errore (c,d)

nei limiti della finestra temporale, in quanto i valori di longevità dell'errore sulla pagina (LE) riportati per gli errori non corretti non corrispondono ad un reale evento di correzione ma alla chiusura dell'intervallo temporale di analisi sperimentale. Analizzando il tempo di persistenza degli errori sulle pagine in base alla tipologia di queste, si nota che per gli errori corretti in pagine *core* il tempo medio di longevità dell'errore sulla pagina è di 0.63 ore

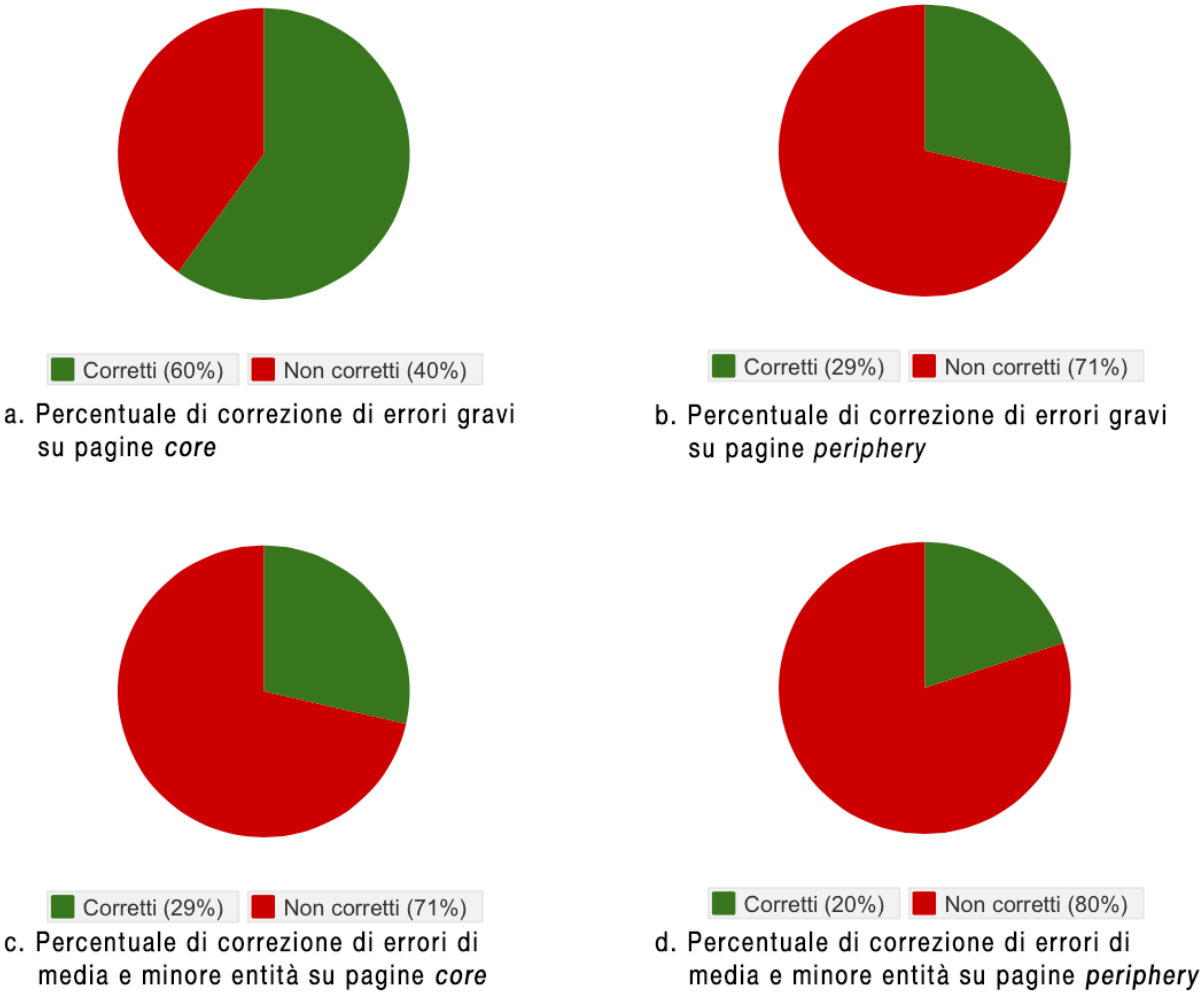


Figura 3.6: Distribuzione delle correzioni suddivise per combinazione di tipologia di pagina e entità di errore

($\sigma = 0.2$) , mentre per gli errori su pagine periphery è di 102 ore ($\sigma = 94$) . L'alto valore di deviazione standard per quest'ultimo gruppo si spiega con il fatto che il calcolo comprende anche le correzioni su 2 pagine periferiche della categoria Calcio, che però possiedono medie di correzione di 1 ora, comparabili con quelle delle pagine *core*; questo è dovuto nel primo caso (*Marcatori dei campionati italiani di calcio*) ad un processo di correzione derivante da

un tracking intra-categoria da parte di un Producer e nel secondo caso al fatto che la pagina (*Campionato mondiale di calcio*), sebbene per numero di legami faccia parte della periferia della rete, ha ricevuto un picco di visualizzazioni nel periodo oggetto dello studio a causa della crescita di interesse da parte degli utenti sull'argomento dovuta all'inizio dei Campionati mondiali di calcio 2014, per cui l'aumento del numero di utenti che ha raggiunto la pagina ha permesso verosimilmente una correzione puntuale dell'errore da parte di un Prosumer in tempi molto brevi. Risultati comparabili sono stati ricavati valutando il tempo di persistenza degli errori, oltre che in base alla pagina in cui sono stati iniettati, anche a seconda della loro entità: per errori gravi inseriti in pagine core, la persistenza media sulla pagina è stata di 0.63 ore ($\sigma = 0.2$); per errori sulle pagine periphery, la persistenza media sulla pagina è stata di 1 ora per errori gravi e di 102 ore per errori medio/piccoli ($\sigma = 28.9$). Questi risultati mostrano la grande stabilità della rete nel nucleo, in cui gli errori vengono corretti con maggiori probabilità e in tempi che sono 2 ordini di grandezza inferiori di quelli associati a errori di entità medio/bassa sulle pagine periferiche; per quanto riguarda queste ultime, si nota come errori gravi siano corretti con tempi comparabili a quelli inseriti nelle pagine core.

Il legame tra correzione dell'errore e tipologia di utente che l'ha eseguita mostra che il 60% degli errori su pagine *periphery* è stato corretto in maniera puntuale da Prosumer, ma in tempi elevati, con una media di persistenza degli errori corretti di 170 ore (fatta eccezione per il caso di errore di grave entità sulla pagina *Campionato mondiale di calcio*). Il restante 40% degli errori su pagine *periphery* è stato corretto mediante revert da Producer ma tramite processi di tracking derivanti dalla correzione di errori su pagine *core* della stessa categoria. Le correzioni da parte di Producer invece state effettuate unicamente su pagine core tramite revert e in tempi molto brevi (meno di 1 ora). Questi risultati, sebbene limitati dallo scarso numero di correzioni registrate durante gli esperimenti, mostrano che le correzioni sulle pagine della periphery sono state effettuate tendenzialmente tramite

accessi puntuali e specialistici da Prosumer, esperti dell'argomento, e non tramite revert generiche come invece avviene da parte dei Producer, comportamenti che rispecchiano le ipotesi dello studio. Per quanto riguarda gli schemi di tracking, non ne è stata rilevata l'esistenza, se non a livello intra-categoria nel caso delle correzioni delle pagine della categoria Calcio, per le quali l'utente Producer ha propagato le sue correzioni partendo dalla pagina più rilevante arrivando fino a quella meno acceduta seguendo le modifiche effettuate dall'utente. Esaminando i dati risultanti dalla registrazione del numero di modifiche su ogni pagina avvenute successivamente all'iniezione dell'errore (nM), si evidenzia la presenza, con frequenza anche in questo caso corrispondente alla rilevanza della pagina all'interno della rete, di modifiche sulle pagine che non portano però alla correzione dell'errore iniettato. Questo comportamento si può descrivere con l'esistenza di un'approccio di tipo additivo per le modifiche che vengono effettuate su Wikipedia; ciò vale per entrambe le tipologie di attori, poichè le modifiche "non-correttive" registrate comprendono sia interventi di tipo "amministrativo" (ad es. correzione di errori ortografici, gestione dei template, formattazione) tipiche dei Producer, che interventi di tipo contenutistico come l'aggiunta di nuove informazioni da parte dei Prosumer.

Considerando infine il numero di visualizzazioni ricevute negli ultimi 30 giorni dalle pagine oggetto dello studio [2] (Figura 3.7), si può notare come la persistenza dell'errore all'interno di pagine appartenenti al core della rete comporti una diffusione dell'errore potenzialmente molto elevata come ci aspetta da reti core-periphery, considerando come indice di diffusione dell'errore il numero di utenti che si collegano alla pagina e visualizzano potenzialmente l'errore. Dai risultati sperimentali si evidenzia come questo comportamento si sia verificato ovviamente per le pagine che non hanno ricevuto correzioni nella finestra temporale, mentre per quelle che sono state corrette si nota come per le pagine core con errori di grave entità, il numero di visualizzazioni sia stato oltre 2 ordini di grandezza inferiore rispetto a

[2]Dati da *Wikipedia article traffic statistics*, `http://stats.grok.se/it/`

quello delle pagine periferiche, a causa di una minor persistenza degli errori [3].

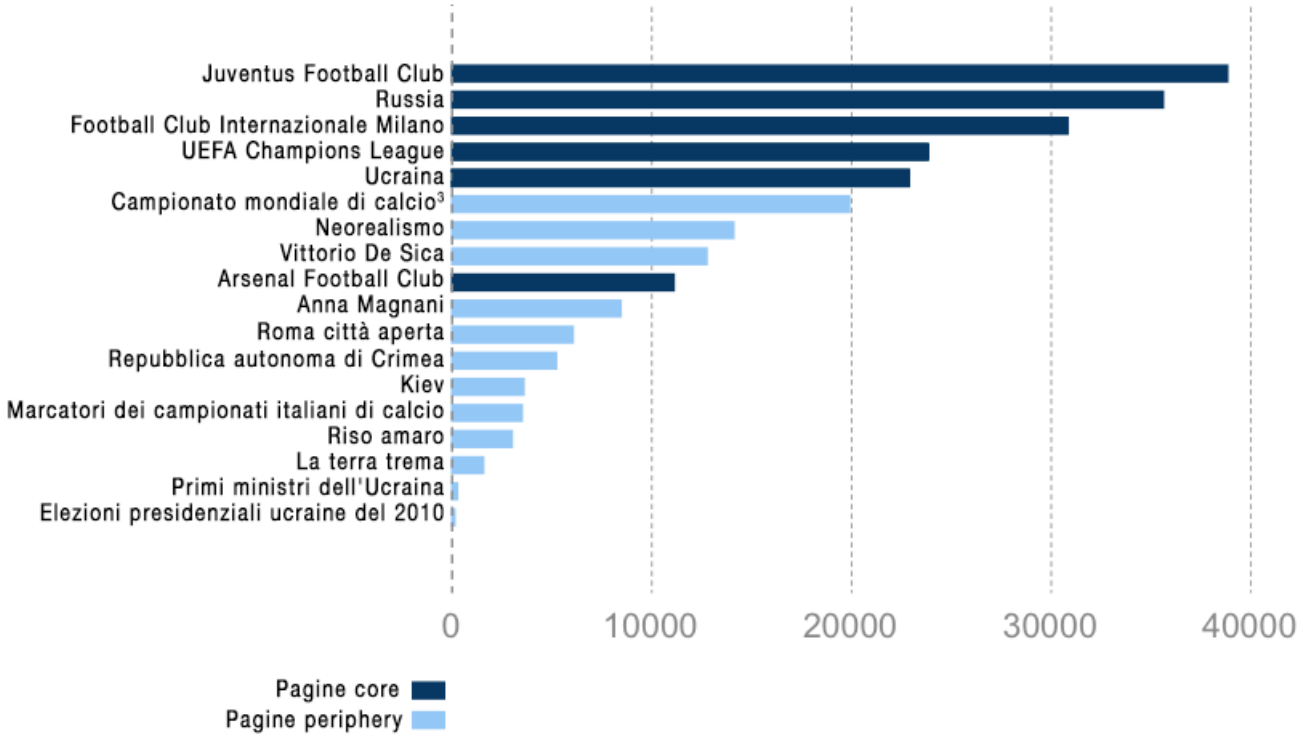


Figura 3.7: Numero medio di visualizzazioni ricevute dalle pagine in 30 giorni

[3]Per la pagina *Campionato mondiale di calcio* il valore di visualizzazioni degli ultimi 30 giorni rappresenta un valore medio basato sulle visualizzazioni degli ultimi 3 mesi. Nella tabella dei risultati il valore nV rappresenta invece l'effettivo numero di visualizzazioni ricevute dalla pagina in presenza dell'errore, valore influenzato dal trend della pagina provocato dall'inizio dei Campionati mondiali di calcio 2014

Il grado di persistenza degli errori nelle pagine può essere evidenziato anche con l'andamento temporale dell'indice rappresentante il numero errori corretti sul totale degli errori, come rappresentato in figura 3.8. Dai grafici si evidenzia una funzione "a gradini" in cui ogni correzione corrisponde uno step verso il basso dell'indice di impatto dell'errore. Nel caso del grafico riferito alla categoria Calcio dell'esperimento 2, due degli step di riduzione dell'indice corrispondono alla chiusura rispettivamente delle fasi 3 e 2 dell'esperimento, per le quali la durata è stata inferiore ai 23 giorni e quindi la chiusura di queste fasi ha portato ad una corrispondente riduzione forzata dell'impatto degli errori sulle pagine. In 3 casi su 6 si nota come l'impatto degli errori sulle pagine persista in maniera costante per tutta la finestra temporale, ciò implica il massimo livello di impatto dell'errore sulle pagine, mentre negli altri casi il valore dell'indice resta comunque elevato, non scendendo mai sotto la soglia dello 0.6, fatta eccezione per la categoria Calcio nell'esperimento 1, per la quale tutte le correzioni sono state eseguite nel primo step temporale. Incrociando questi risultati con il numero medio di visualizzazioni ricevute dalle pagine in 30 giorni (Figura 3.7), si può dedurre come il livello di diffusione degli errori sia effettivamente comparabile con l'andamento del numero di visualizzazioni totali sulle pagine.

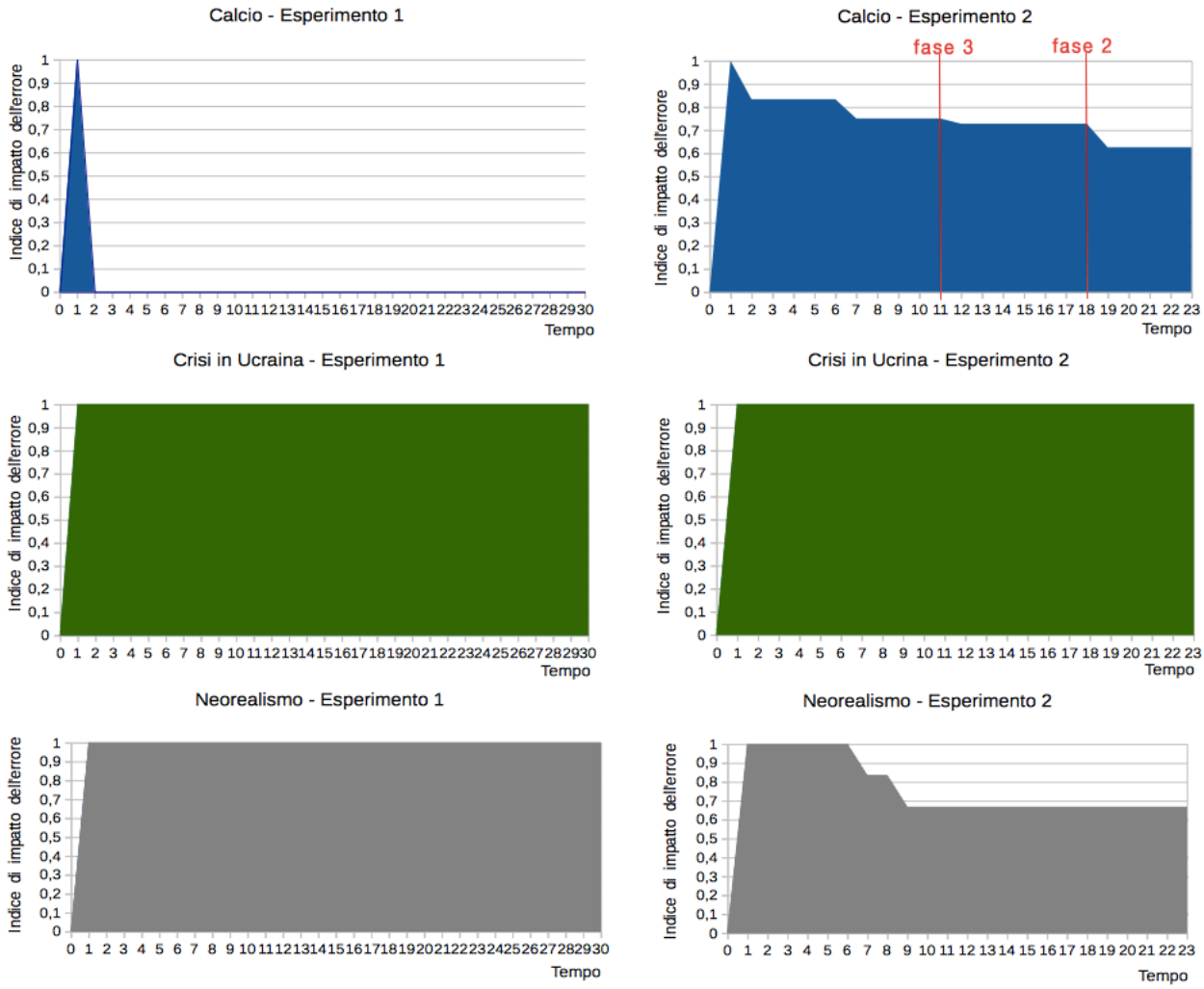


Figura 3.8: Persistenza degli errori nel tempo

Calcio						
Pagina	**Errore**	**LE**	**nV**	**nM**	**TC**	**MC**
Juventus Football Club	Grave	0.4	30	0	S-D	R
Juventus Football Club	Medio	0.4	30	0	S-D	R
Juventus Football Club	Minore	0.4	30	0	S-D	R
Arsenal Football Club	Grave	0.7	15	0	S-T	R
Arsenal Football Club	Medio	0.7	15	0	S-T	R
Arsenal Football Club	Minore	0.7	15	0	S-T	R
Marcatori dei campionati italiani	Grave	1	5	0	S-T	R
Marcatori dei campionati italiani	Medio	1	5	0	S-T	R
Marcatori dei campionati italiani	Minore	1	5	0	S-T	R
Crisi in Ucraina						
Pagina	**Errore**	**LE**	**nV**	**nM**	**TC**	**MC**
Ucraina	Grave	720	22979	25	-	-
Ucraina	Medio	720	22979	25	-	-
Ucraina	Minore	720	22979	25	-	-
Kiev	Grave	720	3697	4	-	-
Kiev	Medio	720	3697	4	-	-
Kiev	Minore	720	3697	4	-	-
Elezioni presidenziali ucraine del 2010	Grave	720	236	0	-	-
Elezioni presidenziali ucraine del 2010	Medio	720	236	0	-	-
Elezioni presidenziali ucraine del 2010	Minore	720	236	0	-	-
Neorealismo						
Pagina	**Errore**	**LE**	**nV**	**nM**	**TC**	**MC**
Vittorio De Sica	Grave	720	12859	3	-	-
Vittorio De Sica	Medio	720	12859	3	-	-
Vittorio De Sica	Minore	720	12859	3	-	-
Roma città apertà	Grave	720	6161	3	-	-
Roma città apertà	Medio	720	6161	3	-	-
Roma città apertà	Minore	720	6161	3	-	-
Riso amaro	Grave	720	3112	1	-	-
Riso amaro	Medio	720	3112	1	-	-
Riso amaro	Minore	720	3112	1	-	-

Tabella 3.4: Tabella riassuntiva delle correzioni dell'esperimento 1

Calcio						
Pagina	**Errore**	**LE**	**nV**	**nM**	**TC**	**MC**
Football Club Internazionale Milano	1	552	23717	39	-	-
Football Club Internazionale Milano	2	552	23717	39	-	-
UEFA Champions League	1	552	18355	8	-	-
UEFA Champions League	2	552	18355	8	-	-
Campionato mondiale di calcio	1	552	41400	121	-	-
Campionato mondiale di calcio	2	159	12000	25	U-D	P
Crisi in Ucraina						
Pagina	**Errore**	**LE**	**nV**	**nM**	**TC**	**MC**
Russia	1	552	27376	20	-	-
Russia	2	552	27376	20	-	-
Repubblica autonoma di Crimea	1	552	4083	4	-	-
Repubblica autonoma di Crimea	2	552	4083	4	-	-
Primi ministri dell'Ucraina	1	552	283	0	-	-
Primi ministri dell'Ucraina	2	552	283	0	-	-
Neorealismo						
Pagina	**Errore**	**LE**	**nV**	**nM**	**TC**	**MC**
Anna Magnani	1	147	1520	0	U-D	P
Anna Magnani	2	552	6560	1	-	-
Neorealismo	1	202	3479	1	U-D	P
Neorealismo	2	552	10890	3	-	-
La terra trema	1	552	1284	0	-	-
La terra trema	2	552	1284	0	-	-

Tabella 3.5: Tabella riassuntiva delle correzioni dell'esperimento 2 - fase 1

Calcio						
Pagina	**Errore**	**LE**	**nV**	**nM**	**TC**	**MC**
Juventus Football Club	1	432	23356	16	-	-
Arsenal Football Club	1	432	6727	3	-	-
Marcatori dei campionati italiani	1	432	2164	0	-	-

Tabella 3.6: Tabella riassuntiva delle correzioni dell'esperimento 2 - fase 2

Calcio						
Pagina	**Errore**	**LE**	**nV**	**nM**	**TC**	**MC**
Football Club Internazionale Milano	1	0.8	50	0	S-D	R
UEFA Champions League	1	264	8778	3	-	-
Campionato mondiale di calcio	1	1	400	0	U-D	P

Tabella 3.7: Tabella riassuntiva delle correzioni dell'esperimento 2 - fase 3

Capitolo 4

Conclusioni

Dalle analisi effettuate sul numero di modifiche e sulle occorrenze comuni tra utenti e pagine è stato possibile mostrare che Wikipedia è caratterizzata da distribuzioni power law con invarianza di scala, in linea con gli altri oggetti del Web, e confermare l'ipotesi che la struttura socio-produttiva dell'enciclopedia libera online sia rappresentabile tramite un modello core-periphery e in particolare che ciò influenza i meccanismi di autocorrezione del sistema.

Le evidenze empiriche dettagliatamente descritte nella sezione 3.4, hanno mostrato innanzi tutto l'inefficienza del sistema autocorrettivo e una generale debolezza del sistema agli attacchi vandalici, poichè solo il 27% degli errori sono stati corretti. É stato evidenziato un generale fallimento dei sistemi di correzione ipotizzati, quello "gerarchico", inteso come la presenza di Producer e di tecniche di patrolling, e quello "comunitario" da parte di Prosumer, che non risultano quindi sufficienti a garantire l'affidabilità di Wikipedia al venir meno del rispetto delle linee guida e dei principi di base da parte degli utenti che inseriscono contenuto. Il caso è quello degli utenti vandali, che non rispettano volutamente queste linee guida e per i quali è necessario un controllo attivo piuttosto che preventivo. I Producer attivi sul sistema non riescono a monitorare le "Ultime modifiche"[1] in maniera precisa e quindi

[1]Pagina speciale Ultime modifiche: `http://it.wikipedia.org/wiki/Speciale:UltimeModifiche`

alcuni errori, specialmente se si tratta di pagine non appartenenti al nucleo della rete o se si tratta di errori di entità medio/bassa, tendono a sfuggire ai controlli, "sommersi" dall'elevato numero di modifiche (oltre 15mila al giorno [2]). Non appaiono inoltre schemi di tracking evidenti che mostrino che le correzioni da parte dei Producer possano avvenire in maniera trasversale tra le differenti categorie. Per quanto riguarda le correzioni degli errori da parte di utenti Prosumer, anche queste appaiono influenzate dalla rilevanza della pagina e l'entità dell'errore: è stato mostrato come le correzioni puntuali da parte di Prosumer che "mantengono" le pagine in quanto esperti dell'argomento vengano effettuate anche su errori di entità medio/bassa ma in tempi elevati, mentre interventi precisi in tempi rapidi sono rilevabili solo per pagine con alta affluenza di utenti e per errori di grande entità. Da parte di entrambe le classi di attori, inoltre, si evidenzia un approccio di tipo additivo per quanto riguarda le modifiche alle pagine, che comporta la permanenza degli errori anche a fronte di interventi sulle pagine.

Per quanto riguarda il costo dell'errore, è stato evidenziato che la struttura di rete core-periphery comporta da un lato i vantaggi di una maggior stabilità del nucleo che implica una maggior probabilità di correzione degli errori inseriti in pagine appartenenti a questo, ma al tempo stesso porta un grande rischio di propagazione dell'errore nel caso in cui questo non venga rilevato e corretto a causa del maggior numero di visualizzazioni ricevute dalle pagine core. La grande stabilità della rete nel nucleo è mostrata anche dai tempi di correzione degli errori, che sono 2 ordini di grandezza inferiori di quelli rilevati per errori di entità medio/bassa sulle pagine periferiche; per quanto riguarda queste ultime, si nota come errori gravi siano corretti con tempi comparabili a quelli inseriti nelle pagine core.

La risposta al perché ci si fida di Wikipedia non sembra quindi debba essere cercata nella sicurezza fornita dall'esistenza di linee guida, regole di correzione o sistemi di patrolling che agiscono attivamente sugli errori, quan-

[2] Statistiche ufficiali Wikimedia `https://stats.wikimedia.org/IT/ChartsWikipediaIT.htm`

to piuttosto nell'approccio dei Wikipediani nei confronti del sistema, che contribuiscono agli articoli dell'enciclopedia per lo spirito di collaborazione e il desiderio di poter far parte di una comunità il cui scopo è la creazione di un sapere comune. In questa ottica però è importante che gli utilizzatori tengano presente, come ricordato anche da Wikipedia stessa, che le informazioni possono essere modificate anche da utenti malevoli. L' apertura a modifiche da parte di qualsiasi utente comporta rischi per i quali non sembra sia possibile sviluppare un sistema correttivo efficace che possa garantire la precisione dell'informazione, soprattutto su pagine marginali.

4.1 Punti deboli

Da un punto di vista dei risultati ottenuti, le debolezze dello studio sono dovute principalmente al numero limitato di pagine utilizzate negli esperimenti a causa di vincoli computazionali. É ipotizzabile che la diffusione degli errori su un numero maggiore di pagine possa portare a risultati più evidenti sopratutto in termini di distribuzione delle correzioni tra le tipologie di utenti e per quanto riguarda le evidenze degli schemi di correzione, che in questo lavoro non è stato possibile analizzare con precisione a causa del basso numero di correzioni registrate. L'utilizzo di un maggior numero di errori sulle pagine estratte in questo studio è stata volutamente evitata per non influenzare i risultati con iniezioni massive, sarebbe quindi necessario applicare le stesse modalità di iniezione degli errori su un campione più ampio di pagine, mantenendo la suddivisione in categorie eventualmente considerando un maggior numero di set indipendenti e un numero maggiore di pagine per ciascuna di essi.

4.2 Sviluppi futuri

Gli elementi di analisi delle reti sociali utilizzati in questo lavoro sono stati fondamentali per la selezione delle pagine e per modellizzare la strut-

tura socio-produttiva di Wikipedia, anche se sono stati limitati allo scopo dello studio. Partendo dai risultati ottenuti si potranno approfondire le dinamiche delle reti con particolare riferimento all'evoluzione temporale che queste hanno avuto dalla nascita di Wikipedia ad oggi. Un'altra interessante prospettiva di studio derivante dai risultati ottenuti in questo lavoro è il confronto del comportamento rilevato per Wikipedia in versione italiana con altre edizioni di Wikipedia, prestando particolare attenzione alle differenze riguardanti articoli scritti originariamente in italiano e articoli che derivano da traduzioni dai corrispondenti in lingua originale.

Appendice A

Prima Appendice

A.1 Listati codice applicativo

A.1.1 Parser per estrazione revision per pagina

EstrazioneRevision.java

```
import java.io.FileNotFoundException;
import java.io.FileOutputStream;
import java.io.FileReader;
import java.io.IOException;
import java.io.PrintStream;
import org.xml.sax.InputSource;
import org.xml.sax.SAXException;
import org.xml.sax.XMLReader;
import org.xml.sax.helpers.XMLReaderFactory;

public class EstrazioneRevision {

public static void main(String[] args) throws SAXException,
FileNotFoundException, IOException {
```

```
try {
FileOutputStream fout = new FileOutputStream("stdout.log");
FileOutputStream ferr = new FileOutputStream("stderr.log");

PrintStream stdout = new PrintStream(fout);
PrintStream stderr = new PrintStream(ferr);

System.setOut(stdout);
System.setErr(stderr);
} catch (FileNotFoundException ex) {
// Could not create/open the file
}

// First pass - to determine headers
XMLReader xr = XMLReaderFactory.createXMLReader();
HeaderHandlerWiki handler = new HeaderHandlerWiki();
xr.setContentHandler(handler);
xr.setErrorHandler(handler);

FileReader r = new FileReader("itwiki-20140508-stub-meta-history.xml");
xr = XMLReaderFactory.createXMLReader();

DataHandlerWiki datahandler = new DataHandlerWiki();

xr.setContentHandler(datahandler);
xr.setErrorHandler(datahandler);

xr.parse(new InputSource(r));
}

}
```

DataHandlerWiki.java

```
import org.xml.sax.Attributes;
import org.xml.sax.helpers.DefaultHandler;

public class DataHandlerWiki extends DefaultHandler {

private String content;
private String currentElement;
private boolean insideElement = false;
private boolean ok = false;
private boolean salta = false;

private boolean prevIdB = false;
private boolean prevIdContribB = false;

private String prevId = null;
private String prevIdContrib = null;

private String titoloPagina = null;
private String idPagina = null;

public DataHandlerWiki() {

super();
}

@Override
public void startElement(String uri, String name, String qName,
Attributes atts) {
currentElement = qName;
content = null;
```

```
insideElement = true;

if ("id".equalsIgnoreCase(qName) && "ns".equalsIgnoreCase(prevId)) {
prevIdB = true;
}
prevId = currentElement;

if ("id".equalsIgnoreCase(qName)
&& "username".equalsIgnoreCase(prevIdContrib)) {
prevIdContribB = true;
}
prevIdContrib = currentElement;

if (qName.equalsIgnoreCase("revision")) {

}

}

@Override
public void endElement(String uri, String name, String qName) {

// se non è l'elemento MEDIAWIKI o l'elemento PAGE
if (!"page".equalsIgnoreCase(qName)
&& !"mediawiki".equalsIgnoreCase(qName)) {

// Se l'elemento in questione è il TITLE con il content che voglio
if ("title".equalsIgnoreCase(qName)) {
if (content != null && qName.equals(currentElement)
&& content.trim().length() > 0
&& (content.equals("Luchino Visconti"))) {
```

```
ok = true;
titoloPagina = content;
// System.out.print(content+",");
} else {
salta = true;
ok = false;
}
}
if (prevIdB && qName.equals("id")) {
idPagina = content;
prevIdB = false;
}

if (ok) {

if (content != null
&& content.trim().length() > 0
&& (qName.equals("contributor")
|| qName.equals("title") || qName.equals("id")
|| qName.equals("timestamp") || qName
.equals("username"))) {

if (!qName.equalsIgnoreCase("contributor")
&& !qName.equalsIgnoreCase("revision")
&& !qName.equalsIgnoreCase("title")) {
System.out.print(content + ",");
}
if (qName.equals("id") && prevIdContribB) {
System.out.print(titoloPagina + "," + idPagina);
prevIdContribB = false;
}
```

```
}

if (qName.equalsIgnoreCase("revision")) {
System.out.println("");
// System.out.print(qName+"_");
}
}
}

// Se è finito l'elemento ITEM
if ("page".equalsIgnoreCase(qName)) {

if (!salta) {

ok = false;

System.out.println("");

}

}

// FINE
insideElement = false;
currentElement = null;
// attribs = null;

}

@Override
```

```
public void characters(char ch[], int start, int length) {
if (insideElement) {
content = new String(ch, start, length);

}
}

public void setHeaderArray(String[] headerArray) {
// this.headerArray = headerArray;
}
}
```

HeaderHandlerWiki.java

```
import java.util.LinkedHashMap;
import java.util.Map.Entry;

import org.xml.sax.Attributes;
import org.xml.sax.helpers.DefaultHandler;

public class HeaderHandlerWiki extends DefaultHandler {

private String content;
private String currentElement;
private boolean insideElement = false;
private Attributes attribs;
private LinkedHashMap<String, Integer> itemHeader;
private LinkedHashMap<String, Integer> accumulativeHeader = new LinkedHashMap<St

public HeaderHandlerWiki() {
super();
}
```

```
LinkedHashMap<String, Integer> getHeaders() {
return accumulativeHeader;
}

private void addItemHeader(String headerName) {
if (itemHeader.containsKey(headerName)) {
itemHeader.put(headerName, itemHeader.get(headerName) + 1);
} else {
itemHeader.put(headerName, 1);
}
}

@Override
public void startElement(String uri, String name, String qName,
Attributes atts) {
if ("page".equalsIgnoreCase(qName)) {
itemHeader = new LinkedHashMap<String, Integer>();
}
currentElement = qName;
content = null;
insideElement = true;
attribs = atts;
}

@Override
public void endElement(String uri, String name, String qName) {
if (!"page".equalsIgnoreCase(qName)
&& !"mediawiki".equalsIgnoreCase(qName)) {
if (content != null && qName.equals(currentElement)
&& content.trim().length() > 0) {
```

```
addItemHeader(qName);
}
if (attribs != null) {
int attsLength = attribs.getLength();
if (attsLength > 0) {
for (int i = 0; i < attsLength; i++) {
String attName = attribs.getLocalName(i);
addItemHeader(attName);
}
}
}
}
if ("page".equalsIgnoreCase(qName)) {
for (Entry<String, Integer> entry : itemHeader.entrySet()) {
String headerName = entry.getKey();
Integer count = entry.getValue();
if (accumulativeHeader.containsKey(headerName)) {
if (count > accumulativeHeader.get(headerName)) {
accumulativeHeader.put(headerName, count);
}
} else {
accumulativeHeader.put(headerName, count);
}
}
}
insideElement = false;
currentElement = null;
attribs = null;
}

@Override
```

```
public void characters(char ch[], int start, int length) {
if (insideElement) {
content = new String(ch, start, length);
}
}
}
```

MultiOutputStream.java

```
import java.io.IOException;
import java.io.OutputStream;

public class MultiOutputStream extends OutputStream
{
OutputStream[] outputStreams;

public MultiOutputStream(OutputStream... outputStreams)
{
this.outputStreams= outputStreams;
}

@Override
public void write(int b) throws IOException
{
for (OutputStream out: outputStreams)
out.write(b);
}

@Override
public void write(byte[] b) throws IOException
{
for (OutputStream out: outputStreams)
```

```
out.write(b);
}

@Override
public void write(byte[] b, int off, int len) throws IOException
{
for (OutputStream out: outputStreams)
out.write(b, off, len);
}

@Override
public void flush() throws IOException
{
for (OutputStream out: outputStreams)
out.flush();
}

@Override
public void close() throws IOException
{
for (OutputStream out: outputStreams)
out.close();
}
}
```

A.1.2 Parser per estrazione dati utente

```
import javax.xml.parsers.DocumentBuilderFactory;
import javax.xml.parsers.DocumentBuilder;
import org.w3c.dom.Document;
import org.w3c.dom.NodeList;
import org.w3c.dom.Node;
```

```
import org.w3c.dom.Element;
import java.io.File;
import java.io.FileNotFoundException;
import java.io.FileOutputStream;
import java.io.PrintStream;

public class ParseDatiUtente {

public static void main(String argv[]) {

try {

String path = "";

FileOutputStream fout = new FileOutputStream("file.csv");
PrintStream stdout = new PrintStream(fout);

System.setOut(stdout);

File folder = new File(path);
File[] listOfFiles = folder.listFiles();
File fXmlFile = null;

System.out
.println("id_user,username,registration_ts,edicount,group");

for (int i = 0; i < listOfFiles.length; i++) {
if (listOfFiles[i].isFile()
&& listOfFiles[i].getName().endsWith("xml")) {

fXmlFile = new File(path + "/" + listOfFiles[i].getName());// MAC
```

```
DocumentBuilderFactory dbFactory = DocumentBuilderFactory
.newInstance();
DocumentBuilder dBuilder = dbFactory.newDocumentBuilder();
Document doc = dBuilder.parse(fXmlFile);

doc.getDocumentElement().normalize();

NodeList nList = doc.getElementsByTagName("user");

for (int temp = 0; temp < nList.getLength(); temp++) {
String utente = "";
Node nNode = nList.item(temp);

if (nNode.getNodeType() == Node.ELEMENT_NODE) {

Element eElement = (Element) nNode;

utente += eElement.getAttribute("userid") + ",";
utente += eElement.getAttribute("name") + ",";
utente += eElement.getAttribute("registration")
+ ",";
utente += eElement.getAttribute("editcount") + ",";

Node nodo = eElement.getElementsByTagName("groups")
.item(0);
if (nodo != null && nodo.hasChildNodes()) {
NodeList nListGruppi = nodo.getChildNodes();
for (int t = 0; t < nListGruppi.getLength(); t++) {
try {
String gruppo = nListGruppi.item(t)
```

```
.getFirstChild()
.getTextContent();
if (!gruppo.equals("*")
&& !gruppo.equals("user")) {
System.out.println(utente
+ nListGruppi.item(t)
.getFirstChild()
.getTextContent());
}
} catch (NullPointerException npe) {
System.err.println(fXmlFile
.getAbsolutePath().toString());
}
}
} else
System.out.println(utente);
}
}
}
}
} catch (FileNotFoundException ex) {
ex.printStackTrace();
} catch (Exception e) {
e.printStackTrace();
}
}
}
```

Appendice B

Seconda Appendice

B.1 Dettaglio degli errori iniettati per ciascuna pagina durante l'esperimento 1

Categoria Calcio - Juventus Football Club

Errore grande

Posizione Primo paragrafo

Frase originale ...stabilmente nella massima categoria del campionato italiano di calcio (dal 1929 denominata Serie A)sin dalla sua fondazione, eccezion fatta per la stagione 2006-07

Modifica ...eccezion fatta per le stagioni **1961-62** e 2006-07

Errore medio

Posizione Sezione Stadio

Frase originale L'opera, progettata dagli studi GAU e Shesa sotto il coordinamento degli architetti Gino Zavanella ed Eloy Suarez e dell'ingegnere Massimo Majowecki

Modifica ...dell'ingegnere Massimo Majowecki **in collaborazione con Giulio Ballio**

Errore piccolo

Posizione Cronologia delle sedi sociali

Frase originale Dal 1923 al 1933: Corso Marsiglia

Modifica Dal 1923 al 1933: Corso Marsiglia **(già via Arcangelo Corelli)**

Categoria Calcio - Arsenal Football Club

Errore grande

Posizione Primo paragrafo

Frase originale ... una delle dodici squadre che hanno raggiunto le finali di tutte le tre principali competizioni gestite dall'UEFA: Champions League (2005-2006), Coppa UEFA (1999-2000) e Coppa delle Coppe (1979-1980, 1993-1994 e 1994-1995)

Modifica ... Coppa delle Coppe (1979-1980, 1993-1994 e 1994-1995) **e al CFU Club Championship (1997-1998)**

Errore medio

Posizione Sezione Storia

Frase originale La rivoluzione tecnico-tattica di Chapman portò l'Arsenal al suo periodo di maggior successo, l'acquisto di giocatori di grande fama come Alex James e Cliff Bastin ...

Modifica ...l'acquisto di giocatori di grande fama come Alex James, Cliff Bastin e **Eric Brook**

Errore piccolo

Posizione Sezione Stadio

Frase originale ... ospitò le gare interne dei Gunners dal settembre 1913 al maggio 2006

Modifica ... dal settembre 1913 al maggio **2008**

Categoria Calcio - Marcatori dei campionati italiani di calcio

Errore grande

Posizione Tabella 1

Frase originale 2010-11 - Antonio di Natale - Udinese

Modifica 2010-11 - Antonio di Natale - **Napoli**

Errore medio

Posizione Vincitori classifica marcatori per squadra

Modifica Invertita la posizione in classifica delle squadre Inter e Juventus

Errore piccolo

Posizione Sezione Note. Nota nr. 2

Frase originale ... i soli due giocatori in grado di segnare un gol in più rispetto alle presenze, rispettivamente 29 reti in 28 presenze ...

Modifica ... rispettivamente **34** reti in **32** presenze ...

Categoria Crisi in Ucraina - Ucraina

Errore grande

Posizione Quinta riga

Frase originale La lingua ufficiale Ë l'ucraino. Molto diffuso nelle regioni orientali e nel sud (in particolare in Crimea) il russo

Modifica ... nel sud (in particolare in Crimea) il russo **e l'osseto**

Errore medio

Posizione Tabella di destra

Frase originale Indipendenza dall'URSS: 25 dicembre 1991 (riconosciuta)

Modifica 25 dicembre **1992**

Errore piccolo

Posizione Sezione Popolazione

Frase originale Il 67% (2005) della popolazione vive in aree urbane

Modifica Il **76**% (2005) della popolazione

Categoria Crisi in Ucraina - Kiev

Errore grande

Posizione Seconda riga

Frase originale Conta 2,6 milioni di abitanti

Modifica Conta **3,6** milioni di abitanti

Errore medio

Posizione Sezione Caratteristiche

Frase originale La città dispone di 3 linee della metropolitana

Modifica ... dispone di **4** linee della metropolitana

Errore piccolo

Posizione Tabella di destra

Frase originale Prefisso: +380 44

Modifica Prefisso: **+390** 44

Categoria Crisi in Ucraina - Elezioni presidenziali Ucraine 2010

Errore grande

Posizione Prima riga

Frase originale ... la vittoria di Viktor Yanukovych, che è divenuto Presidente della Repubblica

Modifica ... che è divenuto **secondo** Presidente della Repubblica

Errore medio

Posizione Tabella di destra

Frase originale Presidente uscente: Viktor Yushchenk

Modifica Presidente uscente: Viktor **Yanukovych**

Errore piccolo

Posizione Prima tabella

Modifica Invertito l'ordine dei candidati al terzo e quarto posto

Categoria Neorealismo - Vittorio De Sica

Errore grande

Posizione Prime righe

Frase originale ... è considerato uno dei padri del Neorealismo e, allo stesso tempo, uno dei più grandi registi ed interpreti della Commedia all'italiana.

Modifica è stato attore di teatro, **divo del muto, primo divo del sonoro**, regista, documentarista, sceneggiatore.

Errore medio

Posizione Sezione Biografia - Attore cinematografico

Frase originale Molto intense anche le sue interpretazioni drammatiche, su tutte quella de Il generale Della Rovere, di Roberto Rossellini (1959), o la partecipazione nel remake di Addio alle armi di Charles Vidor (1957).

Modifica Nella parte finale della propria carriera artistica si trovò ad interpretare ruoli secondari in film anche molto lontani dalla sua immagine, come nel caso di Dracula cerca sangue di vergine... e morì di sete!!!, per la regia di Paul Morrissey (1974), **dove conobbe Andy Warhol**.

Errore piccolo

Posizione Sezione Biografia - gli inizi in teatro

Frase originale Si calcola che De Sica, tra il 1923 e il 1949, abbia preso parte, tra commedie, spettacoli di rivista e drammi in prosa, a oltre 120 rappresentazioni.

Modifica ... a oltre **220** rappresentazioni.

Categoria Neorealismo - Roma città aperta

Errore grande

Posizione Trama

Frase originale ... tradisce l'uomo denunciandolo a Ingrid, agente della Gestapo al servizio del comandante Bergmann

Modifica ... al servizio del comandante **Ingmar** Bergmann

Errore medio

Posizione Tabella di destra

Frase originale Genere: drammatico, guerra

Modifica Genere: drammatico, **guerra civile**

Errore piccolo

Posizione Quarta riga

Frase originale Venne presentato in concorso al festival di Cannes del 1946 dove ottenne il Grand Prix come miglior film

Modifica ... come miglior film **straniero**

Categoria Neorealismo - Riso amaro

Errore grande

Posizione Inizio di paragrafo

Frase originale ... dove aveva presentato Caccia tragica, si trovò nella stazione di Torino ...

Modifica ... dove aveva presentato Caccia tragica**, conoscendo J.L. Godard,** si trovò nella stazione di Torino ...

Errore medio

Posizione Sezione Doppiaggio

Frase originale Ne esiste una versione doppiata in inglese dove la voce della protagonista è di Bettina Dickson

Modifica ... la voce della protagonista è di **una giovanissima Maggie Smith**

Errore piccolo

Posizione Seconda riga

Frase originale ... ha ricevuto una candidatura ai Premi Oscar 1951 per il miglior soggetto

Modifica ... per il miglior soggetto **e miglior sceneggiatura originale**

B.2 Dettaglio degli errori iniettati per ciascuna pagina durante l'esperimento 2

B.2.1 Fase 1

Categoria Calcio

Football Club Internazionale Milano

Errore 1

Posizione Sezione Inno

Frase originale Esistono tuttavia altri due vecchi inni .... Il secondo risale invece al marzo del 1984, Cuore nerazzurro, composto ed eseguito dal gruppo musicale dei Camaleonti.

Modifica ...eseguito dal gruppo musicale dei Camaleonti**, cantato da Richi Maiocchi**

Errore 2

Posizione Sezione Stadio

Frase originale L'impianto, la cui costruzione iniziò nel dicembre 1925 per volere di Piero Pirelli, allora presidente del Milan ...

Modifica per volere di Piero Pirelli **con un progetto degli ingegneri Finzi e Cugini**, allora presidente ...

Categoria Calcio

UEFA Champions League

Errore 1

Posizione Sezione Trofeo

Frase originale ... il trofeo attuale, datato 2005, porta incisi sul retro i nomi di tutte le squadre che l'hanno vinto in precedenza, e ha le orecchie più aggraziate.

Modifica ... ha le orecchie più aggraziate **ed è stata alleggerita a 7.9 Kg**

Errore 2

Posizione Sezione La formula attuale (fine paragrafo)

Frase originale Anche nella finale a turno unico, in caso di parità si giocheranno tempi supplementari e rigori

Modifica ... si giocheranno tempi supplementari **con formula Golden Goal** e rigori

Categoria Calcio

Campionato mondiale di calcio

Errore 1

Posizione Sezione Il Trofeo

Frase originale La nuova coppa è alta 36 cm, fatta di oro 18 carati e pesante 6.175 grammi.

Modifica ... pesante 6.175 grammi **(oltre 1 kg in meno dell'originale)**.

Errore 2

Posizione Sezione Premi

Frase originale Guanto d'oro per il miglior portiere. Viene assegnato dal 1994 e fino al 2006 prendeva il nome di Premio Yashin

Modifica prendeva il nome di Premio Yashin in onore dello storico portiere russo **Alexei** Yashin

Categoria Crisi in Ucraina
Russia

Errore 1

Posizione Sezione Territorio

Frase originale ... le principali isole sono la Novaja Zemlja, la Terra di Francesco Giuseppe, le Isole della Nuova Siberia, líIsola di Wrangel e, sul lato pacifico, le Isole Curili e Sachalin.

Modifica ... e, sul lato pacifico, le Isole Curili, Sachalin, **Kiska e Attu**

Errore 2

Posizione Sezione Demografia

Frase originale I consistenti flussi migratori in uscita (tedeschi di Russia verso la Germania, ebrei verso Israele, russi in cerca di lavoro verso l'Europa occidentale) sono stati parzialmente compensati ...

Modifica ... tedeschi di Russia verso la Germania, ebrei verso Israele, russi in cerca di lavoro verso l'Europa occidentale, **tartari in Cina e Turchia**

Categoria Crisi in Ucraina
Repubblica autonoma di Crimea

Errore 1

Posizione Prima sezione

Frase originale Nel corso della sua storia plurimillenaria, la Crimea ha visto passare sul suo territorio tanti popoli e dominazioni diverse: cimmeri, greci, sciti, sarmati, romani e bizantini, goti, unni, genovesi, tartari, veneziani, turchi, russi e ucraini.

Modifica ... romani e bizantini, goti, unni, **longobardi**, genovesi, tartari, veneziani, turchi ...

Errore 2

Posizione Tabella laterale

Frase originale Motto (Prosperità in unità)

Modifica Motto (Prosperità **e** unità)

Categoria Crisi in Ucraina
Primi ministri dell'Ucraina

Errore 1

Posizione Prima riga

Frase originale Il primo ministro dell'Ucraina è la carica istituzionale che presiede il Governo dell'Ucraina, organo che detiene il potere esecutivo.

Modifica presiede **la Corte Suprema e** il Governo dell'Ucraina

Errore 2

Posizione Paragrafo Autorità

Frase originale ... propone inoltre le candidature dei ministri al Parlamento (eccetto il Ministro degli Affari Esteri e quello della Difesa)

Modifica ... **nomina i 18 giudici della Corte Costituzionale,** propone inoltre le candidature dei ministri al Parlamento (eccetto il Ministro degli Affari Esteri e quello della Difesa)

Categoria Neorealismo

Anna Magnani

Errore 1

Posizione Prima Sezione

Frase originale Celebri le sue interpretazioni, soprattutto in film come Roma città aperta, Bellissima, Mamma Roma e La rosa tatuata

Modifica soprattutto in film come Roma città aperta, Bellissima, **il secondo e terzo episodio di Siamo Donne,** Mamma Roma e La rosa tatuata

Errore 2

Posizione Sezione Il Successo

Frase originale Nel 1947 vince il suo secondo Nastro d'Argento e il premio per la miglior attrice alla Mostra internazionale d'arte cinematografica di Venezia per il film L'onorevole Angelina diretto da Luigi Zampa.

Modifica vince il suo secondo Nastro d'Argento **e la Coppa Volpi**, premio per la miglior attrice alla Mostra internazionale d'arte cinematografica di Venezia,**(vinto prima di lei da un'attrice italiana solo nel 1936, Annabella in Vigilia D'armi)**

Categoria Neorealismo

Neorealismo

Errore 1

Posizione Sezione Il Movimento

Frase originale Il film di Visconti Ossessione, del 1943, è considerato il primo film neorealista ...

Modifica Il film di **Ottone** Visconti ...

Errore 2

Posizione Sezione Il Movimento

Frase originale Molto rilevante è la posizione acquisita dalle riviste, tra cui primeggiava Il Politecnico di Elio Vittorini.

Modifica tra cui **primeggiavano Il Menabò e** il Politecnico di Elio Vittorini

Categoria Neorealismo

La terra trema

Errore 1

Posizione Prima riga

Frase originale La terra trema è un film del 1948 diretto da Luchino Visconti e ispirato al capolavoro del verismo I Malavoglia di Giovanni Verga.

Modifica ... e ispirato al capolavoro del **naturlismo** I Malavoglia ...

Errore 2

Posizione Sezione Trama

Frase originale Aci Trezza, porticciolo vicino ad Acireale. La famiglia Valastro ...

Modifica Aci Trezza, porticciolo vicino ad Acireale, **fine del XVIII secolo**. La famiglia Valastro ...

B.2.2 Fase 2

Categoria Calcio
Juventus Football Club

Errore 1

Posizione Sezione Simboli ufficiali

Frase originale Il nome del club è impresso in caratteri neri e sottolineato in oro su di un'area bianca convessa

Modifica Il nome del club è impresso in caratteri neri **con ombreggiatura dorata** ...

Categoria Calcio
Arsenal Football Club

Errore 1

Posizione Prima sezione, fine secondo paragrafo

Frase originale ... i tifosi dell'Arsenal sono soprannominati Gunners, in italiano "Cannonieri".

Modifica ... i tifosi dell'Arsenal sono soprannominati **Golden** Gunners, in italiano "Cannonieri **Dorati**".

Categoria Calcio
Marcatori dei campionati italiani di calcio

Errore 1

Posizione Primo paragrafo

Frase originale Nel 1929-30 fu istituita la Serie A a girone unico

Modifica Nel 1929-30 fu istituita la Serie A a girone unico **con play-off**

B.2.3 Fase 3

Categoria Calcio
Football Club Internazionale Milano

Errore 1

Posizione Primo paragrafo

Frase originale Fu fondata il 9 marzo 1908

Modifica Fu fondata il 9 marzo **1928 a Seregno**

Categoria Calcio
UEFA Champions League

Errore 1

Posizione Primo paragrafo

Frase originale ... è il più prestigioso torneo internazionale calcistico in Europa per squadre di club maschili.

Modifica ... è il più prestigioso torneo internazionale calcistico in Europa per squadre di club maschili e **femminili**

Categoria Calcio
Campionato mondiale di calcio

Errore 1

Posizione Primo paragrafo

Frase originale ... si disputa ogni 4 anni

Modifica ... si disputa ogni **6** anni

Bibliografia

[1] J. M. Reagle Jr., "In good faith: Wikipedia collaboration and the pursuit of the universal encyclopedia", New York University, Steinhardt School of Culture, Education, and Human Development, 2008

[2] N. Boccara, "Modeling Complex Systems: Second Edition, Graduate Texts in Physics", Springer Science+Business Media, 2010

[3] J. Goodwin "L'autorità di Wikipedia", Sistemi Intelligenti a. XXV, n. 1, Aprile 2013 pp 9-37, 2013

[4] J.N. Cummings, R. Cross, "Structural properties of work groups and their consequences for performance",Social Networks 25, 197-210, 2003

[5] S. Wasserman, K. Faust, "Social Network Analysis: Methods and applications", Structural Analysis in the Social Sciences, Cambridge University Press, Cap. 2;8, 1994

[6] R. Cross *et al.*, "Knowing What We Know: Supporting Knowledge Creation and Sharing in Social Networks", Organizational Dynamics, Vol. 30, No. 2, pp. 100-120, 2001

[7] P. Csermely *et al.*, "Structure and dynamics of core/periphery networks", Journal of Complex Networks 1, 93-123, 2013

[8] R. Priedhorsky, "Creating, Destroying, and Restoring Value in Wikipedia", GROUP-07, November 4-7, 2007, Sanibel Island, Novembre 2007

[9] R. Cross *et al.*, "Knowing What We Know: Supporting Knowledge Creation and Sharing in Social Networks", Organizational Dynamics, Vol. 30, No. 2, pp. 100-120, 2001

[10] P. D. Magnus, "On trusting Wikipedia", EPISTEME, pp. 75-90, 2009

[11] P.D. Magnus, "Early response to false claims in Wikipedia", First Monday, Volume 13 Number 9, Settembre 2008, 2008

[12] F. B Viegas *et al.*, "Studying cooperation and conflict between authors with history flow visualizations", Proc. CHI, 2004

[13] L. Buriol *et al.*, "Temporal Evolution of the Wikigraph", Proc. of the Web Intelligence Conference. Hong Kong, IEEE CS Press, 2006

[14] A. Capocci *et al.*, "Preferential attachment in the growth of social networks: the case of Wikipedia". arXiv.org/physics/0602026, 2006

[15] C. Anderson, "The Long Tail", Wired Magazine, 12-10-2006

[16] W. E. Leland *et al.*, "On the selfsimilar nature of Ethernet traffic", D. P. Sidhu (Ed.), Proc. ACM SIGCOMM, San Francisco, pp. 183-193, 1993

[17] O. Bodin, B.I. Crona, "The role of social networks in natural resource governance: What relational patterns make a difference?", Global Environmental Change 19, 366-374, 2009

[18] M. Deery, L.K. Jago, "The core and the periphery: an examination of the flexible workforce model in the hotel industry",International Journal of Hospitality Management 21, 339-351, 2002

[19] J.N. Cummings, R. Cross, "Structural properties of work groups and their consequences for performance",Social Networks 25, 197-210, 2003

[20] H. Ebel *et al.*, "Scale-free topology of e-mail networks.", Phys. Rev. E, 2002

[21] M. Faloutsos *et al.*, "On power-law relationships of the Internet topology",Computer Comm. Rev. 29 pp. 251-262, 1999

[22] S. Whittaker *et al.*, "The dynamics of mass interaction" ,Proc. CSCW, 1998

[23] P. Gill *et al.*, "Youtube Traffic Characterization: A View from the Edge,Proc. ACM SIGCOMM International Measurement Conference (IMC), San Diego, 2007

[24] A. Mislove *et al.*, "Measurement and Analysis of Online Social Networks", Proc. ACM SIGCOMM International Measurement Conference (IMC), San Diego, 2007

[25] E. H. Chi *et al.*, "Long Tail of user participation in Wikipedia", The Augmented Social Cognition Research Group Blog at Palo Alto Research Center (PARC), 2007

[26] C. Wilson, "The Wisdom of the Chaperones - Digg, Wikipedia, and the myth of Web 2.0 democracy", Slate, 2008

[27] A. Kittur *et al.*, "Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisi', Proc. alt.CHI, 2007

[28] B. Adler, L. Alfaro, "A content-driven reputation system for the Wikipedia", In Proc. WWW, 2007

[29] J. Voss, "Measuring Wikipedia", Proc. Scientometrics and Infometrics, 2005

[30] Jim Giles, "Internet encyclopaedias go head to head", NATURE, 900-01, 2005

[31] Encyclopaedia Britannica, "Fatally Flawed. Refuting the recent study on encyclopedic accuracy by the journal Nature", 2006

[32] E. Randall, "How a raccoon became an aardvark", THE NEW YORKER, http://m.newyorker.com/online/blogs/elements/2014/05/how-a-raccoon-became-an-aardvark.html, Maggio 2014

[33] K. Marks, "Power laws and blogs", http://homepage.mac.com/kevinmarks/powerlaws.html, 2003

[34] E. Cogno, "Wikipedia è l'enciclopedia migliore, dal Regno Unito ai paesi arabi", IL FATTO QUOTIDIANO, http://www.ilfattoquotidiano.it/2012/08/07/wikipedia-migliore-enciclopedia-del-mondo-dal-regno-unito\-ai-paesi-arabi/319733/, 2012

[35] T. Bayer, "New comparative study to re-examine the quality and accuracy of Wikipedia", Wikimedia Foundation Blog, https://blog.wikimedia.org/2011/11/02/new-comparative-study-to-re-examine-the-quality-and\-accuracy-of-wikipedia/, 2011

[36] D. Donato *et al.*, "Large scale properties of the webgraph", EUROPEAN PHYSICAL JOURNAL B, VOL. 38, PP. 239-243, Marzo 2004

[37] Nature, "Encyclopaedia Britannica and Nature: a response", http://www.nature.com/nature/britannica/eb_advert_response_final.pdf, Nature, 2006

[38] J. Wales, "Jimmy Wales talks Wikipedia", http://writingshow.com/?page_id=91, 2005

[39] A. Swartz, "Who writes wikipedia?", http://www.aaronsw.com/weblog/whowriteswikipedia, 2006

[40] Wikipedia, "Page requests per day", http://stats.wikimedia.org/EN/TablesUsagePageRequest.htm, (acceduto il 10 Maggio 2014)

[41] Wikipedia, "History of Wikipedia", http://en.wikipedia.org/wiki/History_of_Wikipedia, (acceduto il 10 Maggio 2014)

[42] Wikipedia, "Livelli di accesso degli utenti", http://it.wikipedia.org/wiki/Wikipedia:Livelli_di_accesso_degli_utenti, (acceduto il 10 Maggio 2014)

[43] Wikipedia, "Elenco permessi per gruppi", http://it.wikipedia.org/wiki/Speciale:ElencoPermessiGruppi, (acceduto il 10 Maggio 2014)

[44] Wikipedia, "Cinque pilastri", http://it.wikipedia.org/wiki/Wikipedia:Cinque_pilastri, (acceduto il 10 Maggio 2014)

[45] Wikipedia, "Avvertenze generali", http://it.wikipedia.org/wiki/Wikipedia:Avvertenze_generali, (acceduto il 15 Maggio 2014)

[46] Wikipedia, "Attendibilità di Wikipedia", http://it.wikipedia.org/wiki/Wikipedia:Attendibilita_di_Wikipedia, (acceduto il 15 Maggio 2014)

[47] Wikipedia, "Verificabilità", http://it.wikipedia.org/wiki/Wikipedia:Verificabilita, (acceduto il 15 Maggio 2014)

[48] Wikipedia, "Gestione del vandalismo", http://it.wikipedia.org/wiki/Wikipedia:Gestione_del_vandalismo, (acceduto il 15 Maggio 2014)

[49] Wikipedia, "Vandalismo", https://it.wikipedia.org/wiki/Wikipedia:Vandalismo, (acceduto il 15 Maggio 2014)

[50] Wikimedia Foundation, "Wikimedia Foundation Guiding Principles", https://wikimediafoundation.org/wiki/Resolution:Wikimedia_Foundation_Guiding_Principles#Freedom_and_open_source, (acceduto il 15 Maggio 2014)

Ringraziamenti

Desidero ringraziare il Prof. Marco Ruffino per i fondamentali contributi nella definizione del disegno sperimentale e nell'analisi delle reti, nonchè per il suo supporto e la sua grande disponibilità durante l'intero periodo di sviluppo dello studio.

Ringrazio inoltre Rocío per le consulenze riguardo l'esposizione dei risultati sperimentali e Nicholas per lo studio e la rappresentazione grafica delle funzioni di interpolazione mediante Gnuplot.