Avaluació de la xarxa social de microblogging descentralitzada Garlanet

Àngel Ollé Blázquez

Director: Joan Manuel Marquès Puig

Universitat Oberta de Catalunya
aollebl@uoc.edu

Resum

Garlanet és una implementació alternativa descentralitzada d'una xarxa social de microblogging tipus Twitter on els recursos computacionals són proporcionats pels mateixos usuaris. Per poder avaluar-la, es necessita realitzar una simulació del comportament que tenen els usuaris a la vida real en xarxes tipus Twitter. En aquest article, s'estudia la investigació duta a terme per poder implantar la simulació dels usuaris i la seva posterior validació. En aquest treball es defineix com generar activitat d'usuari tipus Twitter, així com la implementació d'un sistema que permet generar simulacions de comportament d'usuaris per a Garlanet.

I. Estat de l'art

Garlanet[1] és una implementació d'una xarxa social de microblogging descentralitzada alternativa a Twitter que es centra en la privacitat i on els recursos computacionals són proporcionats pels propis usuaris. Aquests recursos són heterogenis, no dedicats i de menor capacitat en comparació als casos característics dels servidors dedicats. Per aquest fet, es disposarà d'un dinamisme més elevat i una menor qualitat del servei repercutint directament al rendiment, la disponibilitat i a l'experiència de l'usuari. Actualment, Garlanet és un projecte en construcció, sent en Joan Manuel Marquès, un dels impulsors. Es pretén avaluar Garlanet a partir de la seva posada en explotació, pel que sorgeix la necessitat de construir un servei de simulació que pugui avaluar el sistema de manera autònoma.

Per poder avaluar Garlanet, es necessita valorar diferents aspectes de la plataforma que tinguin afectació a les seves característiques principals com a sistema distribuït: rendiment, disponibilitat del servei i experiència de l'usuari. Entre aquests aspectes podem trobar: el dinamisme (fallada o desconnexions), la connectivitat i els comportaments dels usuaris de la xarxa que puguin condicionar un o més aspectes de la plataforma.

El procés de recerca d'estudis realitzats amb anterioritat s'ha centrat en dos aspectes: l'activitat dels usuaris i el comportament dels repositoris no dedicats aportats voluntàriament.

Els aspectes que s'han avaluat en altres plataformes descentralitzades semblants com Hornet [2], donen una primera proposta de què s'ha de tenir en compte per contrastar el rendiment. En primer lloc, es pot realitzar una estimació de l'ús de les dades que circulen per la plataforma, aquestes es poden tenir diferents orígens en funció de la font que les genera: els usuaris i la lògica de l'aplicació.

Els arguments que menciona la proposta de Hornet són molt extrapolables a Garlanet degut que parteixen d'una base similar en quant l'aportació voluntària de recursos i la seva finalitat principal. Així mateix, plataformes com BOINC [3] també contribueixen a ampliar les propostes

proporcionant coneixement de com es comporten els recursos no dedicats aportats voluntàriament, sent un símil en finalitat de com ho permet l'arquitectura de Garlanet.

Per tant, és pot afirmar que un dels aspectes més destacables per avaluar Garlanet és el comportament dels seus usuaris. Realitzar una simulació amb els comportaments d'interès que siguin significatius per la plataforma, com l'enviament de missatges, relacions entre ells, tendències globals a partir de comportaments individuals, i altres paràmetres que permetin caracteritzar l'activitat dels usuaris extrets d'estudis cercats.

Amb l'informe de Barracudalabs [4], s'extreuen les estadístiques més importants que són rellevants en l'activitat dels usuaris. Quan es crea un usuari que forma part de la simulació, s'ha de decidir a quin grup formarà part, a partir de les dades extretes de l'informe es realitza la classifiació segons el nombre de *followers* i *following*; aquests valors incideixen a tendir en un comportament en el nombre de missatges enviats.

També, existeixen els estudis de predicció de personalitat dels usuaris [5] que, tot i no ser aplicables directament, es pot aprofitar per extreure característiques implícites, resultant en el que es coneixen com a relacions entre usuaris: usuaris que són seguits per molts usuaris i usuaris que segueixen a molts usuaris.

Les relacions de seguiment entre usuaris, es poden interpretar com a grafs dirigits etiquetats [6, 7, 8, 9] on l'atribut relacional de cada usuari pot ser implementat en aquesta estructura de dades.

Atès que els estudis cercats ofereixen unes característiques útils per a la implementació de la simulació dels usuaris, els aspectes que es consideraran per realitzar la implementació són:

- Nombre de missatges enviats.
- Nombre de seguidors.
- Nombre d'usuaris que segueixen un usuari.
- Distribució de missatges enviats i rebuts en un període de temps.
- Distribució de l'activitat durant un temps acotat.

En conseqüència, a partir de les dades obtingudes en el procés de recerca, es fa una implementació de simulació dels usuaris, modelant el seu comportament que puguin tenir a la vida real a partir de les estadístiques aconseguides i extretes de Twitter.

II. Modelat de Twitter

Per modelar el comportament dels usuaris s'ha seguit les estadístiques recollides de l'estudi "Barracuda Labs 2010 Annual Security Report"[3].

S'han relacionat la quantitat de *followers*, *following* i nombre de missatges. Tots els aspectes estan correlacionats entre ells i s'han de tenir en consideració: el nombre de missatges d'un usuari està relacionat en funció de qui segueix i qui el segueix. A més, s'han d'establir les relacions i proporcions adients entre: *following*, *followers* i quantitat de missatges.

I. Quantitat de followers

La *Figura 1* mostra la quantitat de followers per cada 100 usuaris de Twitter, segons l'informe de *Barracuda Labs*:

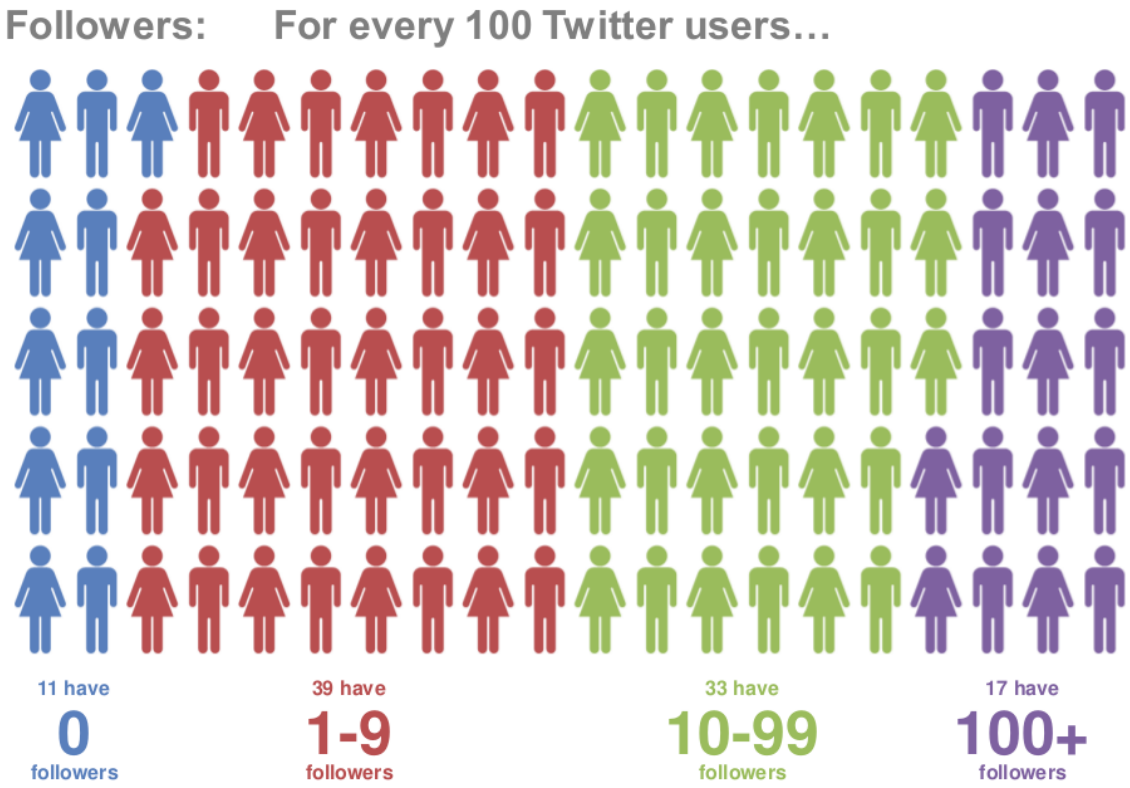

Figura 1: Followers. Barracuda Labs 2010 Annual Security Report

Durant el modelatge, es necessita classificar els *followers* segons aquestes dades. Per decidir a quin grup pertany cada usuari, es fa de la manera següent:

1.- Assignar l'usuari en una categoria de nombre de *followers*.

Es genera un nombre aleatori utilitzant una distribució uniforme discreta [10] en el rang [0,100). Segons el nombre aleatori generat, l'usuari s'assignarà a una de les categories de la *Taula 1*:

Nombre aleatòri	Categoria de followers
0 - 10	0
11 - 49	1 - 9
50 - 82	19 - 99
83 - 99	100+

Taula 1. Categoría de followers

2.- Definir el nombre concret de followers de l'usuari.

Per calcular el nombre exacte de followers que tindrà l'usuari dintre del rang esmentat, generarem el nombre exacte de followers a partir de la *Taula 2*:

Categoria de followers	Nombre de followers
0	0
1 - 9	nombre distrib. uniforme discreta en rang[1,9]
10 - 99	nombre distrib. uniforme discreta en rang[10,99]
100+	nombre (>100) distrib. geomètrica amb alpha 0.005

Taula 2. Nombre de followers

II. Quantitat de following

S'ha de decidir cada usuari a qui segueix per tal que cada usuari tingui els followers que s'han calculat en el punt anterior. Per tal de fer una simulació el més realista possible, el nombre d'usuaris als quals segueix un usuari s'ha d'ajustar a la distribució de la *Figura 2*:

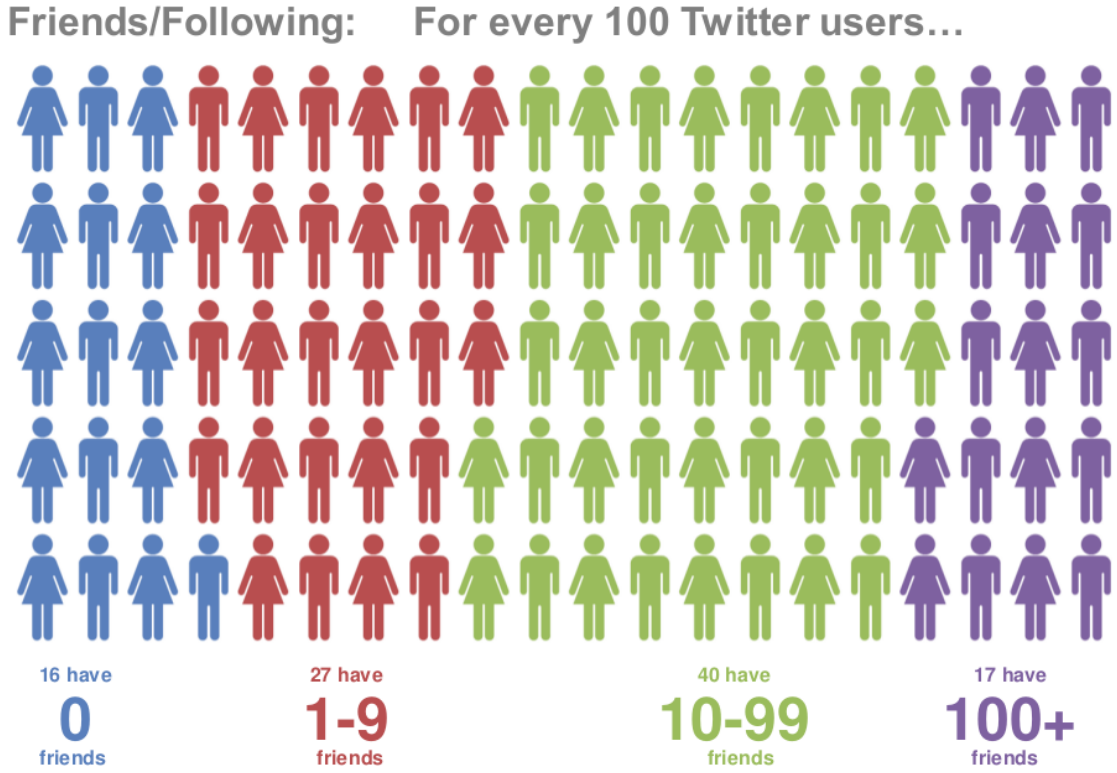

Figura 2. Following. Barracuda Labs 2010 Annual Security Report

Aquest pas s'ha realitzat de manera posterior al càlcul dels *followers* però immediatament abans d'establir les relacions, s'han definit els quatre blocs amb les quantitats màximes per following. En el moment de repartir tots els *followers*, es respecten els valors màxims per bloc de *following* per a cada assignació.

III. Generació d'activitat

Per decidir la quantitat d'activitat que generarà cada usuari, partim de la mitjana diària de missatges enviats pels usuaris en un dia, la *Figura 3* ens ho mostra:

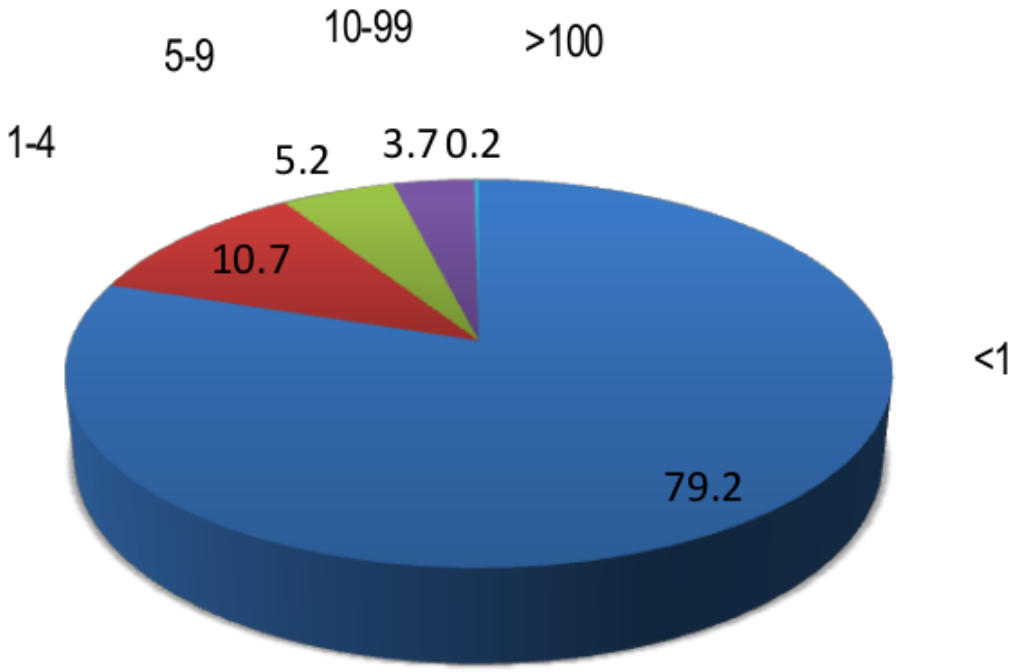

Figura 3. Average Tweets Per Day. Barracuda Labs 2010 Annual Security Report

L'estudi de Barracuda Labs que utilitzem com a base per a simular el comportament dels usuaris no proporciona dades per a poder definir el nivell d'activitat de cada usuari segons el nivell de followers que té. Per aquest motiu hem definit unes taules que permetin definir el nivell d'activitat dels usuaris segons la categoria de followers a la que s'ha assignat l'usuari. La *Taula 3* ens mostra la relació entre nivell d'activitat i categoria de followers sobre el nombre total d'usuaris:

	Categoria followers estudi Barracuda					
	0	**1 – 9**	**10 – 99**	**100+**		
Percentatge d'usuaris estudi Barracuda	11%	39%	33%	17%		
79,4%	10,8%	36,9%	24,9%	6,8%	<1	**Categoria nivells d'activitat (nombre de missatges/dia) estudi Barracuda**
10,9%	0,2%	1,2%	4,5%	5%	1 – 4	
5,4%	0%	0,4%	2%	3%	5 – 9	
3,9%	0%	0,4%	1,5%	2%	10 – 99	
0,4%	0%	0,1%	0,1%	0,2%	100+	

Taula 3. Relació nivell d'activitat segons categoria de followers sobre el nombre total d'usuaris. Technical report. [12]

Cada percentatge de color blau estableix quin percentatge d'usuaris hi haurà en cada combinació de categoria de followers i categoria d'activitat d'usuaris. A continuació es presenta la mateixa informació però, en lloc d'indicar el percentatge d'usuaris sobre el nombre total d'usuaris, es mostra el percentatge d'usuaris de cada categoria de nivell d'activitat per cada categoria de followers. Segons la *Taula 4*:

		Categoria followers			
		0	**1 – 9**	**10 – 99**	**100+**
Categoria nivells d'activitat (nombre de missatges/dia) estudi Barracuda	**<1**	98,18%	94,62%	75,45%	40%
	1 – 4	1,82%	3,08%	13,64%	29,41%
	5 – 9	0%	1,03%	6,06%	17,65%
	10 – 99	0%	1,03%	4,55%	11,76%
	>100	0%	0,26%	0,3%	1,18%

Taula 4. Relació nivell d'activitat segons categoria de followers per cada categoria de followers. Technical report.

Els passos per establir el nombre de missatges que enviarà l'usuari per dia es fan de la manera següent:

1. Assignar usuari en una categoria de nivell d'activitat. A partir de la generació d'un nombre aleatori amb una distribució uniforme entre [0,100). Amb la darrera *Taula 5* de relació de nivell d'activitat segons categoria de followers per cada categoria de followers i el nombre aleatori.

		Categoria de followers			
		0	**1 – 9**	**10 – 99**	**100+**
Categoria nivells d'activitat (nombre de missatges/dia)	**<1**	$0 \leq x < 98{,}18$	$0 \leq x < 94{,}62$	$0 \leq x < 75{,}45$	$0 \leq x < 40$
	1 – 4	$98{,}18 \leq x < 100$	$94{,}62 \leq x < 97{,}71$	$75{,}45 \leq x < 89{,}09$	$40 \leq x < 69{,}41$
	5 – 9	N/A	$97{,}71 \leq x < 98{,}74$	$89{,}09 \leq x < 95{,}15$	$69{,}41 \leq x < 87{,}06$
	10 – 99	N/A	$98{,}74 \leq x < 99{,}77$	$95{,}15 \leq x < 99{,}7$	$87{,}06 \leq x < 98{,}82$
	>100	N/A	$99{,}77 \leq x < 100$	$99{,}71 \leq x < 100$	$98{,}83 \leq x < 100$

Taula 5. Relació nivell d'activitat segons categoria de followers per cada categoria de followers i nombre aleatori. Technical report.

2. Determinació nivell d'activitat diària de cada usuari. Un cop classificat l'usuari, es determinarà el nombre de missatges que enviarà diàriament:

2.1 En el cas que el nombre de missatges sigui inferior a un, tindrem dos casos diferents segons el nivell d'activitat i a partir d'un nombre aleatori [0,100) amb una distribució uniforme discreta:

- Cas d'activitat normal: Si aquest nombre és entre [0, 10], llavors s'enviarà un missatge; en cas contrari, no s'enviarà cap missatge.
- Cas de molta activitat: Si el nombre és entre [0, 50], llavors s'enviarà un missatge; en cas contrari, no s'enviarà cap missatge.

2.2 Per la resta de casos:

- Categoria 1-4: es genera un nombre aleatori [1, 4] amb una distribució uniforme discreta i s'envien la quantitat de missatges que indiqui el nombre.
- Categoria 5-9: es genera un nombre aleatori [5, 9] amb una distribució uniforme discreta i s'envien la quantitat de missatges que indiqui el nombre.
- Categoria 10-99: es genera un nombre aleatori [10, 99] amb una distribució uniforme discreta i s'envien la quantitat de missatges que indiqui el nombre.

- Categoria +100: es genera un nombre amb una distribució geomètrica [11] amb alpha 0.04 i s'envien la quantitat de missatges que indiqui el nombre.

III. Arquitectura

L'arquitectura dissenyada per realitzar la simulació és compon del següent:

ActivitySimulator: conté tota la lògica extreta a partir de l'estudi del funcionament de Twitter. S'encarrega de fer tots els càlculs per treure els 'description' (que està format per: following, followers, nombre de missatges, nom de l'usuari, password, keystore, certificats, etc.) dels usuaris. Serveix els description als usuaris simulats (UserSimulator).

GarlanetSimulator: és el servei que s'encarrega de simular Garlanet, proporciona la capacitat de fer els registres dels usuaris, el login, logout, i l'enviament de missatges als followers dels usuaris.

La darrera part del sistema, **UserSimulator**, representa els clients i és l'encarregat de simular el comportament dels usuaris. Es connecta als dos serveis anteriors, al primer per obtenir el 'description' i al segon per enregistrar-se, fer *login*, enviar missatges i sortir.

La *Figura 4* mostra la representació de l'arquitectura especificada:

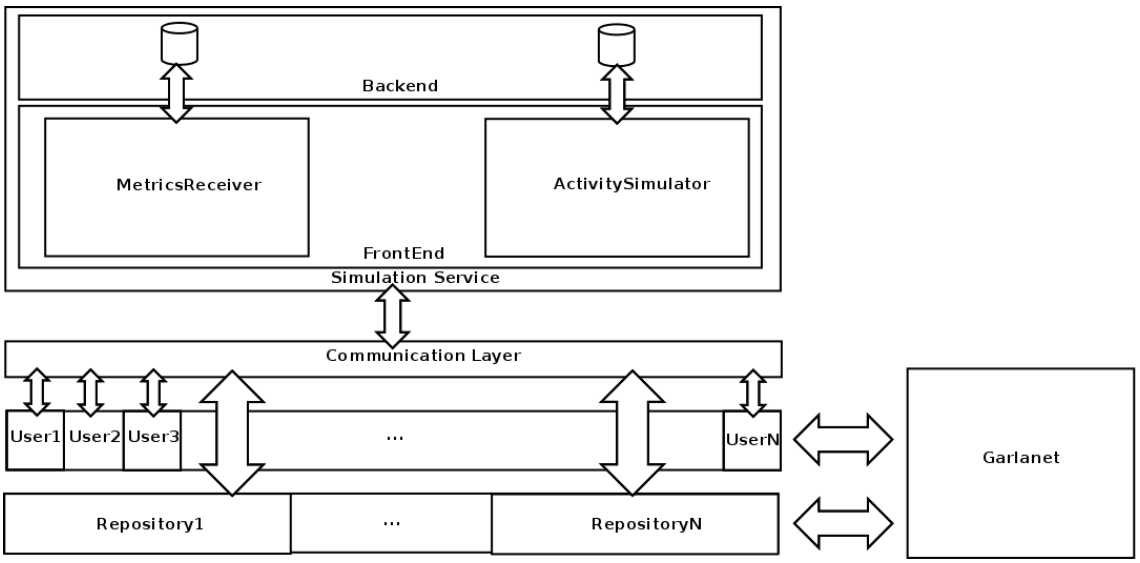

Figura 4. Disseny de l'arquitectura. Technical report.

El flux de les aplicacions (veure *Figura 5*) i la interacció dels usuaris és el següent:

1. L'usuari simulat demana a l'ActivitySimulator participar en l'experiment. Aquest, el registra.
2. l'ActivitySimulator li envia el Description de l'usuari.
3. L'usuari simulat, actua segons el Description rebut: registrant-se, fent login, enviant missatges i fent logout.
4. Durant la interacció amb GarlanetSimulator, l'usuari rep els missatges dels qui segueix.
5. Durant l'experiment, l'usuari envia les mètriques al servei MetricsReceiver, que desa aquestes a una base de dades sqlite.

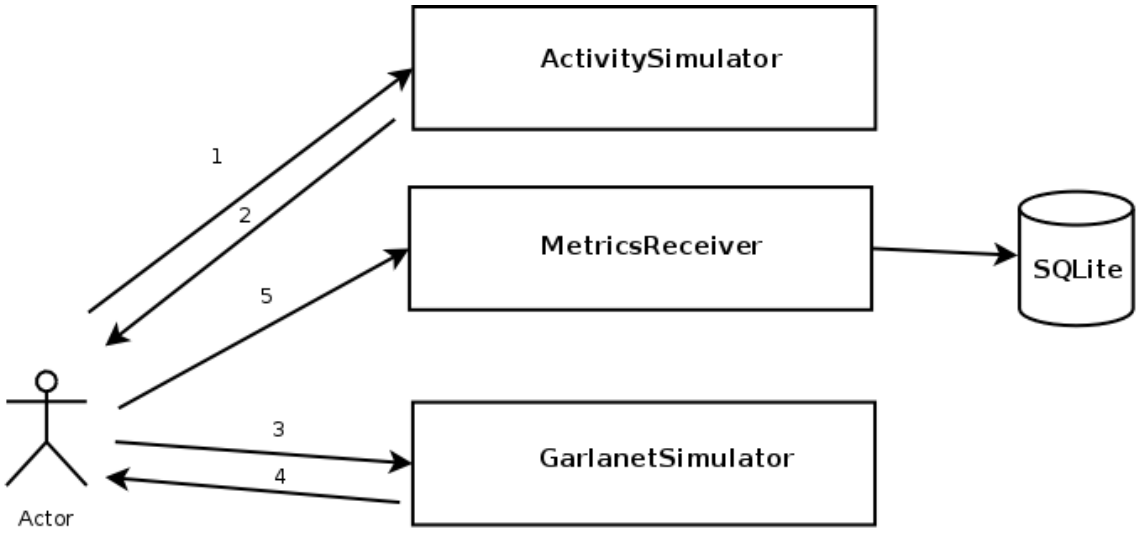

Figura 5. Flux de les aplicacions i interacció amb l'usuari. Technical report.

IV. Implementació

El procés de desenvolupament utilitzat durant l'implementació ha estat el ***Test-driven development (aka TDD)***. Les parts implementades han estat: ActivitySimulator, GarlanetSimulator i l'UserSimulator. S'enumeren les tecnologies i característiques utilitzades més destacables:

- Llenguatge de programació: Java. Versió de JDK: Oracle 1.7.
- Comunicació de xarxa: Netty 4.1. Framework NIO client-servidor.
- Multithreading.
- Rendiment: ús de pautes d'optimització del rendiment.
- Java Generics: ús dels genèrics.
- Proves unitàries: JUnit, Mockito i PowerMock.
- Patrons de disseny.
- Gestió del temps: permet realitzar simulacions en temps real o en temps de simulació.
- Control de versions: Durant el desenvolupament del projecte, s'ha fet servir Git per al manteniment del codi.
- Gestió de dependències, configuració i empaquetament: Maven.
- Shell scripting.
- Base de dades: sqlite3.
- Google Charts.

Per més detall sobre la implementació veure el document *Technical Report*.

V. Validació del model

Per a la validació del model s'han realitzat tres experiments amb parametritzacions diferents. La quantitat d'usuaris durant l'experiment ha estat limitada per les limitacions físiques computacionals des d'on han estat executats.

Se'n detallen a continuació els resultats obtinguts i la comparació amb les dades extretes del *Barracuda Labs 2010 Annual Security Report*. En el cas el nombre de Tweets enviats per dia, notar que en mencionat informe el còmput total no és exacte.

I. Experiment I

El temps de simulació ha estat d'un dia, amb una quantitat de dos-cents usuaris a la xarxa. S'han enviat un total de 821 missatges entre els usuaris.

Quantitat de followers (*Figura 6*)

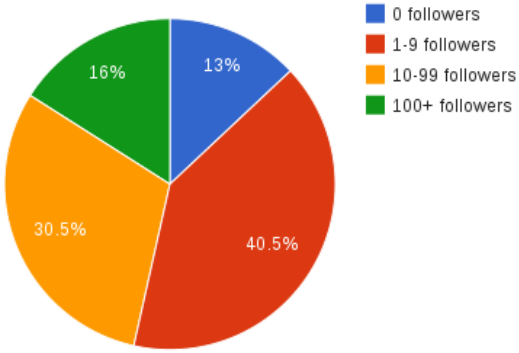

Figura 6. Experiment I. Quantitat de followers.

La comparació dels resultats amb les dades de *Barracuda* es mostra a la *Taula 6*:

Quantitat de followers		
	Exp. I	*Barracuda*
0	13%	11%
1- 9	40.5%	39%
10 - 99	30.5%	33%
100+	16%	17%

Taula 6. Experiment I. Comparació de quantitat de followers

Mitjana de Tweets per dia (*Figura 7*)

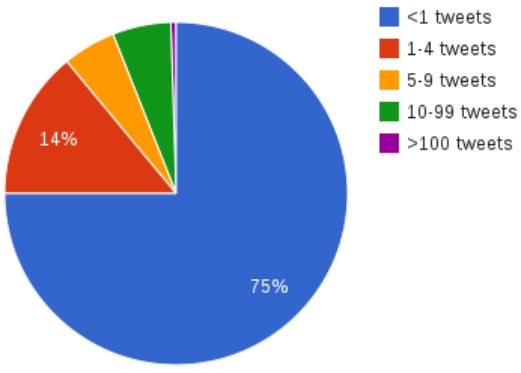

Figura 7. Experiment I. Mitjana de Tweets per dia.

La comparació dels resultats amb les dades de *Barracuda* es mostren a la *Taula 7*:

Mitjana de Tweets per dia		
	Exp. I	*Barracuda*
<1	75%	79.2%
1- 4	14%	10.7%
5 - 9	5%	5.2%
10 - 99	5.5%	3.7%
>100	0.5%	0.2%

Taula 7. Experiment I. Comparació de la mitjana de Tweets per dia

II. Experiment II

El temps de simulació ha estat d'un dia, amb una quantitat de dos-cents usuaris a la xarxa. S'han enviat un total de 552 missatges entre els usuaris.

Quantitat de followers (*Figura 8*)

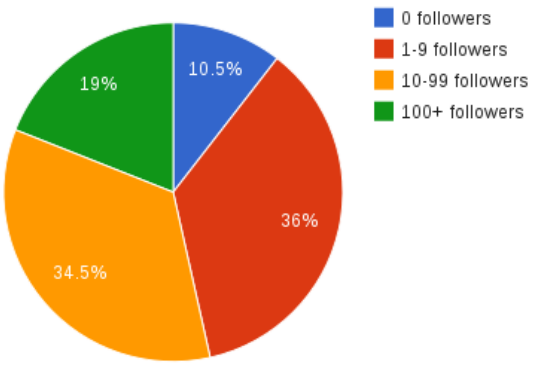

Figura 8. Experiment II. Quantitat de followers.

La comparació dels resultats amb les dades de *Barracuda* es mostra a la *Taula 8*:

Quantitat de followers		
	Exp. II	*Barracuda*
0	10.5%	11%
1- 9	36%	39%
10 - 99	34.5%	33%
100+	19%	17%

Taula 8. Experiment II. Comparació de quantitat de followers

Quantitat de following (*Figura 9*)

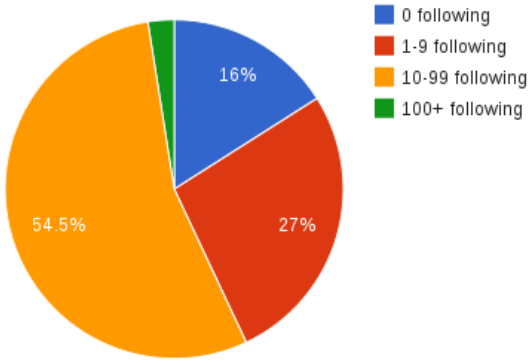

Figura 9. Experiment II. Quantitat de following.

La comparació dels resultats amb les dades de *Barracuda* es mostra a la *Taula 9*:

Quantitat de following		
	Exp. II	*Barracuda*
0	16%	16%
1- 9	27%	27%
10 - 99	54.5%	40%
100+	2,5%	17%

Taula 9. Experiment II. Comparació de quantitat de following

Mitjana de Tweets per dia (*Figura 10*)

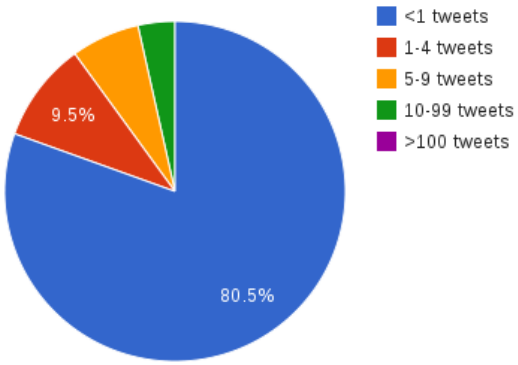

Figura 10. Experiment II. Mitjana de Tweets per dia.

La comparació dels resultats amb les dades de *Barracuda* es mostra a la *Taula 10*:

Mitjana de Tweets per dia		
	Exp. II	*Barracuda*
<1	80.5%	79.2%
1- 4	9.5%	10.7%
5 - 9	6.5%	5.2%
10 - 99	3.5%	3.7%
>100	0%	0.2%

Taula 10. Experiment II. Comparació de la mitjana de Tweets per dia

III. Experiment III

El temps de simulació ha estat de dos dies, amb una quantitat de dos-cents cinquanta usuaris a la xarxa. S'han enviat un total de 1348 missatges entre els usuaris.

Quantitat de followers (*Figura 11*)

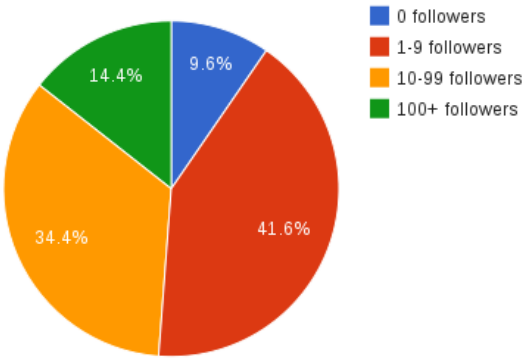

Figura 10. Experiment III. Quantitat de followers.

La comparació dels resultats amb les dades de *Barracuda* es mostra a la *Taula 11*:

Quantitat de followers		
	Exp. III	*Barracuda*
0	9.6%	11%
1- 9	41.6%	39%
10 - 99	34.4%	33%
100+	14.4%	17%

Taula 11. Experiment III. Comparació de quantitat de followers

Quantitat de following (*Figura 12*)

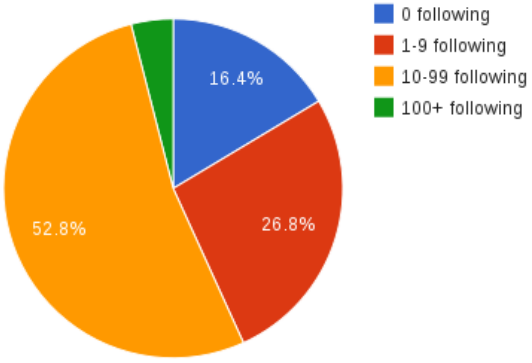

Figura 11. Experiment III. Quantitat de following.

La comparació dels resultats amb les dades de *Barracuda* es mostra a la *Taula 12*:

Quantitat de following		
	Exp. III	*Barracuda*
0	16.4%	16%
1- 9	26.8%	27%
10 - 99	52.8%	40%
100+	4%	17%

Taula 12. Experiment III. Comparació de quantitat de following

Mitjana de Tweets per dia (*Figura 13*)

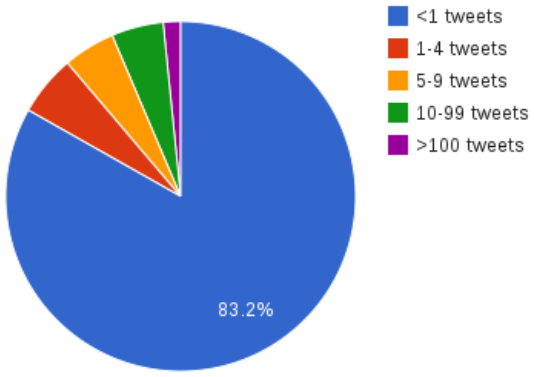

Figura 13. Experiment III. Mitjana de Tweets per dia.

La comparació dels resultats amb les dades de *Barracuda* es mostra a la *Taula 13*:

Mitjana de Tweets per dia		
	Exp. III	*Barracuda*
<1	83.2%	79.2%
1- 4	5.6%	10.7%
5 - 9	4.8%	5.2%
10 - 99	4.8%	3.7%
>100	1.6%	0.2%

Taula 13. Experiment III. Comparació de la mitjana de Tweets per dia

Amb el resultat dels experiments, es pot apreciar la gran semblança de valors comparats amb els de *Barracuda*. Notar que el volum d'usuaris de l'experiment ha estat molt menor als que té realment Twitter i que poden ser les causants de les petites variacions.

VI. Treball futur

Es plantegen els aspectes següents com a possibles millores i treball futur:

- Modelar el comportament dels repositoris igual que s'ha realitzat pel comportament dels usuaris.
- Executar i analitzar experiments amb una quantitat d'usuaris molt més elevada.
- Utilitzar l'ActivitySimulator i l'UserSimulator directament amb Garlanet.
- Aplicar optimitzacions de rendiment.
- Utilitzar l'ActivitySimulator en diferents testbeds (p.ex. planetlab[13], communityLab[14]) i avaluar-ne el comportament.

Referències

[1] Joan Manuel Marquès, Helena Rifà-Pous, *Garlanet, a decentralized microblogging social network*, Technical report UOC2013/07-2, Jul. 2013.

[2] D. Lazaro, J.M. Marques, G. Cabrera, E. Rifa-Pous, A. Montane, *HorNet: Microblogging for a Contributory Social Network*, IEEE Internet Computing, vol.16 N.3 Pg.37-44, Jun. 2012.

[3] BOINC, *Berkeley Open Infrastructure for Network Computing*, http://boinc.berkeley.edu/, 2002-2015.

[4] Barracuda Labs, *Barracuda Labs 2010 Annual Security Report*, barracudalabs.com, 2010.

[5] D. Quercia, M. Kosinski, D. Stillwell, J. Crowcroft, *Our Twitter Profiles, Our Selves: Predicting Personality with Twitter*, Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference, 2011.

[6] T. Yamashita, H. Sato, S. Oyama, M. Kurihara, *Classification of Twitter Users Based on Following Relations*, International MultiConference of Engineers and Computer Scientists 2013, 2013.

[7] P. Pena, R. del Hoyo, J. Vea-Murguia, C. Gonzalez, S. Mayo, *Collective Knowledge Ontology User Profiling for Twitter*, 2th IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2013.

[8] S. Tramp, P. Frischmuth, T. Ermilov, S. Shekarpour, S. Auer, *An Architecture of a Distributed Semantic Social Network*, IOS Press, 2012.

[9] K. Tao, F. Abel, Q. Gao, G. Houben, *TUMS: Twitter-based User Modeling Service*, 8th Extended Semantic Web Conference (ESWC 2011), 2011.

[10] Wolfram, *Uniform Distribution*, mathworld.wolfram.com/UniformDistribution.html, 2015.

[11] Wolfram, *Geometric Distribution*, mathworld.wolfram.com/GeometricDistribution.html, 2015.

[12] Àngel Ollé Blázquez, Joan Manuel Marquès, *TFM: Avaluació de la xarxa social de microblogging descentralitzada Garlanet - Technical Report* , Universitat Oberta de Catalunya, 2015.

[13] PlanetLab, *PlanetLab: An open platform for developing, deploying, and accessing planetary-scale services*, planet-lab.org, 2015.

[14] Community Lab, *Community Lab: Community Networks Testbed by the CONFINE project*, community-lab.net, 2015.