Simulating the emergence of critical mass in online communities: How forums scale by stimulating user contributions

Wybo Wiersma
Oxford Internet Institute
`mail@wybowiersma.net`

1 Introduction

Thread-based discussions are enormously popular on the internet. Not only are they found in usenet-groups, and on classical web-forums, but they also appear in mailing-lists, google-groups, on many news-sites, below blog-posts, and more recently, on Facebook walls. Over decades several attempts have been made at developing alternative discussion-platforms, but apart from Twitter and chat, these have not been successful.[8,43,14,67,68] Is there something special about thread-based discussion-platforms that makes them so good at attaining critical mass?

Are they successful because forum-threads resemble the temporal order of off-line conversations? Is it because people are used to this format on the web? Or is it because they are good at overcoming the critical mass problem? Probably a mixture of these, but here the latter is assumed: it is because forums are good at attaining critical mass. Thus making the research-question: what is it in the structure of threaded forums that allows them to quickly ignite a community into being, and then sustain it as it grows?

An agent-based model, grounded in theory and observation, will be built to clarify and formalize our understanding of this process. The model will be calibrated, and tested, against the structure of real forums of varying sizes. This model will then be used to predict what scale-ranges successful forums can have, how the psychological utility derived from them varies, and how thread-structures change as forums grow. In addition, the model will be used to examine what would happen if certain features of threaded discussions had been missing or were reversed.

The paper proceeds as follows: First some previous work will be discussed. Then some limits of the proposal will be set out. Next, the research-question will be further specified, as well as some of the other questions that will be answered in the thesis. Following that, the research-design will be introduced, and especially its method: agent-based modeling. Then the data and sampling strategy will be presented, followed by a brief discussion of

some challenges that might arise during the project. Finally, the paper will be concluded with some ideas for further research.

1.1 Previous works

Most similar are a set of forthcoming papers by Yuking and Kraut, which introduce and evaluate an agent-based model of user-behaviour on a forum. Their use-case is a test of the influence of moderation-regimes (no moderation, community-wide moderation, and personalized moderation) on user-contributions, and community ties.[56,55] A number of other works employ agent-based modelling for studying how critical mass develops across networks, how it impacts the valuation of software companies, and the diffusion of innovation across various types of networks.[33,7,42].

A notable work that deals with discussion-threads, is Sandra Gonzalez-Bailon's paper about the shape of threads across topics on Slashdot. What she found, for example, is that political threads tended to be much wider and deeper than those about games.[20]. A lot of work has been done on individuals motivations for contributing to on-line communities.[1,2,6,31,39,44,45,46,50,51,52,53,54,57]. The social aspects of threaded discussions have been studied as well, for example by generating social

networks based on reply-structures, and either using them to detect differing roles, or to analyse forums as social communities.[12,34,35,62,66]

Works from the theoretical realm are, first of all those by Peter Hedstrom on analytical sociology, where the focus is on so called social mechanisms. Social mechanisms operate on the micro-level of decisions made by individuals, and bring about macro-level effects through emergence.[26,25,27] A second set of works, are those in Structuration Theory (not to be confused with structuralism), which state that besides individual agency, social, and other structures, are important as well.[15,47] Finally, for grounding the model, and in addition to the listed empirical studies, several more general psychological and economic theories will be used, such as Karau and Williams 'collective effort' model, Rogers Agarwala-Rogers 'information overload' theory, and possibly theories of group-identity, and signalling.[70,30,32,59,29,13,41]

1.2 Limits

First of all, the thesis will be limited to one aspect of virtual communities; the thread-based discussion. Other important aspects of online communities, such as their governance, visual appearance, and other features, such as friending, will not be examined or discussed.

The focus will be on the internal dynamics of forums. External effects, such as competition between communities or advertising-campaigns, will not be taken into account.

Secondly, instead of taking threads as indicative of social structures, threads, and the (social) mechanisms involved in them, will be considered the primary drivers. The assumption here is that most people simply reply to interesting posts they happen to come across, largely ignoring whether they know the author of the post they reply to. If true, then directly social factors may be spurious, especially if structural factors can explain most of the dynamics as well. Also, for reasons of space, the content of posts, their topics, or (emotional) polarity, will not be looked at. This though they are very important for determining which specific threads/posts receive replies, and even for the flourishing of whole communities.[6,21]

In addition we will only use data from two forums as found on the web, so no surveys, interviews about user-motivations, or other elicited data will be used. In the narrow sense, no new data was thus created for this study. Yet in the broader sense of data gathering (such as scraping), we did collect some new data. Finally as this study limits itself to two data-sets, there will be limited generalizability (more on data and generalizability in sec-

tion 5).

2 Questions

The research question, as noted, is what is it in thread-structures that helps forums overcome the critical mass problem. This is an explanatory research question that could lead us to various mechanisms. An example of the kind of mechanism we hope to test is, as most of us probably observed in practice, that first posts generally receive more attention than posts later in the thread. Which, in short, should provide a heightened incentive to early contributors, and thus make people more likely to be the first to post on a (sub) forum. Another example is that, by listing new threads most prominently, forums focus attention on a few topics, increasing the likelihood of interactions.

In order to examine these, and other social mechanisms, we will take the following steps, answering the listed questions along the way. They are ordered according to A.D. de Groots empirical circle, of observation, induction, deduction, testing and evaluation.[22] They each match to a different research-approach.[48]

1. What variations in forum-sizes are actually observed? How small, and how big, do forums get? How does thread-creation- and posting-behaviour vary be-

tween forums of different sizes? Supposedly larger forums will have a greater churn of both threads and posts, but exactly how much faster is this turnover? And does it vary with the time of day? (one of the questions studied for the OSN essay) The goal of this step is to examine and present the descriptive statistics of our data (descriptive approach), which the model should later explain.

2. What social factors / mechanisms should we induce as being potentially important? In what ways do people derive utility from participating in online forums? (studied for formative DEG essay) What is the 'production function' of forums? How much more are earlier posts valued than later ones? What costs, such as time and information-overload, are involved? Exploring the possible mechanisms is what we do here (exploratory).

3. How can we deduce an agent-based model from the above-mentioned factors?: A model with decision rules for how users, their incentives and actions relate. We will here also defend why we should use modelling as a method, and ground the model in theory. The aim here is to formalize our hypotheses and their interactions in the model (constructive).

4. Is the model valid? Here we will test the model against previously unexamined data, and report on the extent to which it fits. We will try to use all the power of the data here. So we will not merely match the model output against real growth figures, but also test for similarly structured threads, and reply-structures. The aim of this question is to find out if the model provides an appropriate covering law (empirical).

5. Finally, what can we conclude from the model? How does the perceived utility of (early) posting change as forums grow? How big can forums get, and how does this match with the sizes that were observed? Is there an optimum size? Finally, how does the model hold against counter-factuals, and what does this teach us about the properties of forums that make them better at gaining critical mass?; such as showing users threads at random, instead of the most recent ones, or appending new posts at the front of the thread? And are there possible improvements, such as showing new threads to new users first? The aim of this step is to clarify and explain the relationships between the factors (explanatory).

4

3   Research design

Our primary research-design will thus be one that is inspired by the empirical tradition. Yet as noted, at the centre of it will be a constructive approach (common in computer science), which means the creation of a solution for a problem: in our case an agent-based model that matches our problem domain. Therefore the fit with the empirical approach is not complete, and arguably we will be employing a mixed method design.[9].

4   Modelling

In sociology the use of agent-based modelling has been pioneered by Nigel Gilbert, Thomas Schelling, Dirk Helbing, and several others. They have applied it to domains such as innovation networks, racial segregation, and pedestrian behaviour in crowded streets.[17,60,28] The method has thus been tried and tested with some success, even though it is relatively new.

On a more theoretical level the core tenet of agent-based modelling is taking the point of view of individuals, and modelling their behaviour: for example based on their expected utility.[5] Sociological processes only come into the model by emerging from the behaviour of individual agents. This approach matches that of Peter Hedstrom, who is known for propos-

ing that group behaviour is best understood as emerging from social mechanisms operating at the level of individuals.[26,25,27]

Apart from this theoretical foundation, and its practical success, agent-based modelling has other advantages as well. The first being that it allows one to integrate multiple, specific, smaller theories. Where in most of the social sciences (particularly in experimental psychology) one normally tests only one variable, while keeping the others constant, agent-based modelling allows one to combine such theories into a middle-level framework.[16]

Another, closely coupled advantage, is that the model/framework is not just written down in a narrative (as it previously would have been), but rather is formalized in the code of the model. This both enables it to be more strictly defined, and allows us to test the model against hard data.[4] These properties are what makes the method especially useful for helping us clarify and understand complex, emergent phenomena, such as the appearance of critical mass in web-forums.

A final advantage is that, contrary to equation-based modelling of relationships between variables or aggregates (which would have been the alternative approach) agent-based modelling allows one to open the black-box of micro processes. This can be done because these (in the form of the agents decision

rules) form the building-blocks of the model. And it is also this modularity, which allows us to use agent-based models as test-beds, when experimenting with counter-factuals.

For agent-based modelling to be done successfully, it is important that the model be well-grounded in theory, and is plausible. This is because a model that matches the data is not necessarily correct. Multiple models resulting in the same growth-patterns/output can always be created, without any way to decide between them, other than Occams razor or perhaps taste. While this is not a new problem (any empirical study trying to establish causality suffers from it), but the fact that we work with multiple independent variables, some of which could even be hypothetical, instead of measured, makes the need for theoretical grounding more pressing.

We will build our agent-based model using NetLogo, an Open Source platform for agent-based modelling. A first draft of the model has already been formulated.

5   Data

Because not much research has been done on web-forums, and data-sets are relatively labour intensive to create/pre-process, our sampling strategy is not based on a random selection of web-forums/news-sites, etc. In-stead we picked two typical, yet different sites as our case studies. The first source, and 90% of the second, will be used for informing and tweaking the model. The remaining 10% of the latter will then be used to validate it.

The first data-set is the Hacker News news-site, which is comparable to Slashdot, in that people can post links to news stories, and comment on them.[61,23] It is a bit smaller than Slashdot, though still quite large, with 90.000 unique visitors per day, most of whom are lurkers. What makes Hacker News especially suitable for our study, is that all threads and posts on it are rated by users, which will give us a quantitative measure of the value of earlier posts relative to later ones. Data has been collected from the site for three weeks now, and this will continue until the end of April.

The second data-set is 30 gigabytes of data from Boards.ie, the largest Irish bulletin board site.[3] This site offers hundreds of web-forums on various topics. The two things that make this data-set especially useful, are, first of all, that the forums for the various topics differ in size (allowing us to establish the relationship between community size and posting behaviour/thread structure), and secondly, that the data is longitudinal (1998 to 2008), showing how each sub-forum was created and gained momentum over time. This data-set was collected and provided by John Breslin,

from NUI Galway.[10]

6 Challenges

There are several challenges that have to be overcome for this thesis-project. First of all, constructing the model, and especially making it match well with the data, will be the most labour intensive part of the thesis. We will especially have to guard against spending too much time polishing the model. As it could literally be improved endlessly. In addition, answering all the sub-questions within adequate time, might also prove to be a challenge. However, selectively dropping some of the less central ones when needed, will allow for some leeway.

Secondly, agent-based modelling is a relatively new method, and though it has been shown to be successful, convincing some readers of the soundness of it, might be hard. Communicative challenges could arise from the formality, and/or over-complexity of the model. If too many factors are added to the model, it might become incomprehensible even to willing readers.[24] Lastly, there is the grounding in theory, that needs to be done properly. Keeping the model relatively simple, and close to participants intuitions (where possible), should alleviate most of these issues.

Finally, in the ethical realm, there should be no problems. Only documents that are publicly available on the web, are analysed, and then only their thread-structures, not their content. And even though our data contains user-id's, these are mostly nicknames. Also, no specific individuals are being studied, but only average behaviour. Thus, even though the model will consist of agents, these will not be modelled after specific persons, or even be differentiated (e.g. they will all behave according to the same logic). Finally, none of the issues or behaviour studied are sensitive in any way.[11]

The only way in which our study might adversely impact the communities under study, is by changing them: either by making them realise they are public spaces that can be studied, or by making their participants aware of the dynamics of such groups. However, for boards.ie, previous studies have done this already, and besides, this issue is relatively minor, and applies to virtually all research in the social sciences.

7 Conclusion

To conclude, some limits of the proposal were set, and the most important previous work was discussed, including theories that might be used to ground our approach and model.

Next the research-question, and the steps involved, were clarified. The research-design was then specified in detail, giving the rationale for using agent-based modelling. Finally the data-sets were described, and some minor challenges and ethical issues related to the research were discussed.

In all this thesis-project will be an exciting endeavour, that should (tentatively) open up some new ground. At the very least the output of the process should be 1) a simple working (but not perfect) model of user contributions to forums, 2) a clearer picture of the mathematical relationships between community size, and activity volume/types/network-structure (possibly even a set of covering laws), and at least one interrelated set of possible mid-level theoretical explanations of what makes threaded forums so good at overcoming the critical mass problem (formalized in the model).

There are many possibilities for further research. More possible mechanisms could be identified, for example by surveying, interviewing or recording people that regularly interact with forums. The model could be further improved and tweaked, adding topics and emotional polarity as important determiners. The model could also be made more general, and extended to other domains; such as blogs, or even the journal-based publica-

tion system. Naturally, the model can also be tested, and cross-validated against other data-sets. Finally, the model can be used to inform the design of, and (if extended) even indicate the possible success of alternative discussion-platforms.

Bibliography

1. Beenen, G. *et al.*, 'Using social psychology to motivate contributions to online communities.', *Proceedings of the 2004 ACM conference on Computer supported cooperative work* 212–221 (2004).

2. Bimber, B., Flanagin, A. J, & Stohl, C., 'Reconceptualizing collective action in the contemporary media environment.', *Communication Theory* 15, 365–388 (2005).

3. *boards.ie: Now ye're talkin'.* <http://boards.ie/>.

4. Boero, R., & Squazzoni, F., 'Does empirical embeddedness matter? Methodological issues on agent-based models for analytical social science.', *Journal of Artificial Societies and Social Simulation* 8, 6 (2005).

5. Bonabeau, E., 'Agent-based modeling: Methods and techniques for simulating human systems.', *Proceedings of the Na-*

tional Academy of Sciences of the United States of America 99, 7280 (2002).

6. Chen, G., & Chiu, M. M, 'Online discussion processes: Effects of earlier messages' evaluations, knowledge content, social cues and personal information on later messages.', *Computers & Education* 50, 678–692 (2008).

7. Choi, H., Kim, S. H, & Lee, J., 'Role of network structure and network effects in diffusion of innovations.', *Industrial Marketing Management* 39, 170–177 (2010).

8. *Cohere: Make the Connection.* `<http://cohere.open.ac.uk/>`.

9. Creswell, J. W, & Clark, V. L., *Designing and conducting mixed methods research.* (Sage Publications, Inc, 2007).

10. *data.sioc-project.org | Home of the boards.ie SIOC Data Competition.* `<http://data.sioc-project.org/>`.

11. Eynon, R., Fry, J., & Schroeder, R., 'The ethics of internet research.', *The SAGE handbook of online research methods* 23 (2008).

12. Fisher, D., Smith, M., & Welser, H. T, 'You are who you talk to: Detecting roles in usenet newsgroups.', *System Sciences, 2006. HICSS'06. Proceedings of the 39th*

Annual Hawaii International Conference on 3, 59b (2006).

13. Friedman, E. J, & Resnick, P., 'The social cost of cheap pseudonyms.', *Journal of Economics & Management Strategy* 10, 173–199 (2001).

14. Garrett, L. N., Smith, K. E., & Meyrowitz, N., 'Intermedia: issues, strategies, and tactics in the design of a hypermedia document system.', *Proceedings of the 1986 ACM conference on Computer-supported cooperative work* 163–174 (1986).

15. Giddens, A., *The constitution of society: Outline of the theory of structuration.* (University of California press, 1984).

16. Gilbert, G. N, *Agent-based models.* (Sage Publications, Inc, 2008).

17. Gilbert, N., Pyka, A., & Ahrweiler, P., 'Innovation networks-a simulation approach.', *Journal of Artificial Societies and Social Simulation* 4, 1–13 (2001).

18. Goldenberg, J., Libai, B., & Muller, E., 'The chilling effects of network externalities.', *International Journal of Research in Marketing* (2009).

19. Gonzalez-Bailon, S., 'Networks and mechanisms of interdependence: Theoretical developments beyond the rational

action model.', *Revista Internacional de Sociologia* 67 (2008).

20. Gonzalez-Bailon, S., Kaltenbrunner, A., & Banchs, R. E, 'The structure of political discussion networks: A model for the analysis of online deliberation.', *Journal of Information Technology* 25, 230–243 (2010).

21. Gonzalez-Bailon, S., 'The positive effects of negative emotions in online forums.', *Forthcoming* (2011).

22. Groot, A. D, *Methodology: Foundations of inference and research in the behavioral sciences.* (Mouton, 1969).

23. *Hacker news.* <http : / / news . ycombinator.com/>.

24. Hamill, L., 'Agent-based modelling: The next 15 years.', *Journal of Artificial Societies and Social Simulation* 13, 7 (2010).

25. Hedstrom, P., 'Explaining the growth patterns of social movements.', *Understanding Choice, Explaining Behaviour* (2006).

26. Hedstrom, P., & MyiLibrary, *Dissecting the social: on the principles of analytical sociology.* (Cambridge University Press Cambridge, UK, 2005).

27. Hedstrom, P., & Ylikoski, P., 'Causal mechanisms in the social sciences.', *Annual Review of Sociology* 36, 49–67 (2010).

28. Helbing, D., & Molnar, P., 'Social force model for pedestrian dynamics.', *Physical review E* 51, 4282–4286 (1995).

29. Hogg, M. A, & Terry, D. J, 'Social identity and self-categorization processes in organizational contexts.', *Academy of Management Review* 25, 121–140 (2000).

30. Johnson, S. L, & Faraj, S., 'Preferential attachment and mutuality in electronic knowledge networks.', *26th International Conference on Information Systems* 287–299 (2005).

31. Joyce, E., & Kraut, R. E, 'Predicting continued participation in newsgroups.', *Journal of Computer-Mediated Communication* 11, 723–747 (2006).

32. Karau, S. J, & Williams, K. D, 'Social loafing: A meta-analytic review and theoretical integration.', *Journal of Personality and Social Psychology* 65, 681 (1993).

33. Kemper, A., *Valuation of network effects in software markets: A complex networks approach.* (Physica-Verlag Heidelberg, 2009).

34. Kiesler, S., *Culture of the Internet.* (Lawrence Erlbaum, 1997).

35. Kunegis, J., Lommatzsch, A., & Bauckhage, C., 'The slashdot zoo: Mining a so-

cial network with negative edges.', *Proceedings of the 18th international conference on World wide web* 741–750 (2009).

36. Lampe, C., & Johnston, E., 'Follow the (slash) dot: Effects of feedback on new members in an online community.', *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work* 11–20 (2005).

37. Lampe, C., & Resnick, P., 'Slash (dot) and burn: Distributed moderation in a large online conversation space.', *Proceedings of the SIGCHI conference on Human factors in computing systems* 543–550 (2004).

38. Lang, K., 'Newsweeder: Learning to filter netnews.', *in Proceedings of the 12th International Machine Learning Conference (ML95* (1995).

39. Lazar, J., & Preece, J., 'Social considerations in online communities: Usability, sociability, and success factors.', *Cognition in a digital world* 127–151 (2002).

40. Liebowitz, S. J., *Network externalities.* `<http://www.utdallas.edu/~liebowit/palgrave/network.html>`.

41. Ma, M., & Agarwal, R., 'Through a glass darkly: Information technology design, identity verification, and knowledge contribution in online communities.', *Information Systems Research* 18, 42 (2007).

42. Montanari, A., & Saberi, A., 'The spread of innovations in social networks.', *Proceedings of the National Academy of Sciences* 107, 20196 (2010).

43. Nelson, T. H., *Literary machines : the report on and of, project Xanadu, concerning word processing, electronic publishing, hypertext, thinkertoys ...* (1992).

44. Nonnecke, B., & Preece, J., 'Lurker demographics: Counting the silent.', *Proceedings of the SIGCHI conference on Human factors in computing systems* 73–80 (2000).

45. Nonnecke, B., & Preece, J., 'Why lurkers lurk.', *Americas Conference on Information Systems, Boston* (2001).

46. Onnela, J. P, & Reed-Tsochas, F., 'The spontaneous emergence of social influence in online systems.', *Arxiv preprint arXiv:0912.0045* (2009).

47. Orlikowski, W. J, 'The duality of technology: Rethinking the concept of technology in organizations.', *Organization science* 3, 398–427 (1992).

48. Outhwaite, W., & Turner, S. P, *The SAGE handbook of social science methodology.* (SAGE Publications Ltd, 2007).

49. Platt, J., 'Social traps.', *American Psychologist* 28, 641–651 (1973).

50. Prasarnphanich, P., & Wagner, C., 'Creating critical mass in collaboration systems: Insights from Wikipedia.', *2nd IEEE International Conference on Digital Ecosystems and Technologies, 2008. DEST 2008* 126–130 (2008).

51. Prasarnphanich, P., & Wagner, C., 'Explaining the sustainability of digital ecosystems based on the wiki model through critical mass theory.', (2009).

52. Preece, J., Nonnecke, B., & Andrews, D., 'The top five reasons for lurking: Improving community experiences for everyone.', *Computers in Human Behavior* 20, 201–223 (2004).

53. Rafaeli, S., & LaRose, R. J, 'Electronic bulletin boards and "public goods" explanations of collaborative mass media.', *Communication Research* 20, 277 (1993).

54. Ren, Y., Kraut, R., Kiesler, S., *et al.*, 'Applying common identity and bond theory to design of online communities.', *Organization studies* 28, 377–408 (2007).

55. Ren, Y., & Kraut, R. E, 'A simulation for designing online community: Member motivation, contribution, and discussion moderation.', *Information Systems Research* (2011).

56. Ren, Y., & Kraut, R. E, 'Agent-based modeling to inform online community theory and design: Impact of discussion moderation on member commitment and contribution.', (2011).

57. Ren, Y. *et al.*, 'Increasing commitment to online communities: Designing from theory.', *Management Information Systems Quarterly* (2010).

58. Roberts, T. L, 'Are newsgroups virtual communities?', *Proceedings of the SIGCHI conference on Human factors in computing systems* 360–367 (1998).

59. Rogers, E. M, & Agarwala-Rogers, R., 'Organizational communication.', *Communication and behaviour* 218–236 (1975).

60. Schelling, T. C, 'Dynamic models of segregation.', *The Journal of Mathematical Sociology* 1, 143–186 (1971).

61. *Slashdot: News for nerds, stuff that matters.* <http://slashdot.org/>.

62. Smith, M. A, 'Netscan: Measuring and mapping the social structure of usenet.', (1997).

63. *Update on Google Wave - official Google blog.* <http://googleblog.blogspot.com/2010/08/update-on-google-wave.html>.

64. Varian, H. R, 'Intermediate microeconomics.', (2006).

65. Varian, H. R, Farrell, J., & Shapiro, C., *The economics of information technology: An introduction.* (Cambridge Univ Pr, 2004).

66. Welser, H. T, Gleave, E., Fisher, D., & Smith, M., 'Visualizing the signatures of social roles in online discussion groups.', *Journal of Social Structure* 8 (2007).

67. Wiersma, W., & Sarlo, B., 'LogiLogi: a webplatform for philosophers.', *Digital Humanities 2008 Book of Abstracts* 221–222 (2008).

68. Wiersma, W., 'LogiLogi: The quest for critical mass.', *Unpublished thesis* (2010).

69. Williams, K., & Harkins, S., 'Identifiability as a deterrent to social loafing: Two cheering experiments.', *Journal of Personality and Social Psychology* 40, 303–311.

70. Zandt, T. V., 'Information overload in a network of targeted communication.', *RAND Journal of Economics* 35, 542–560 (2004).

71. Zittrain, J., 'The rise and fall of sysopdom.', *Harvard Journal of Law & Technology* 10, 495 (1996).