Supplementary Information for: *Structural limitations of learning in a crowd: communication vulnerability and information diffusion in MOOCs*

Nabeel Gillani [1], Taha Yasseri [2], Rebecca Eynon* [2], Isis Hjorth [2]

[1]
Department of Engineering Science
University of Oxford
Parks Road
Oxford OX1 3PJ
UK
nabeel@robots.ox.ac.uk

[2]
Oxford Internet Institute
University of Oxford
1 St. Giles'
Oxford OX1 3JS
UK
{taha.yasseri, rebecca.eynon, isis.hjorth}@oii.ox.ac.uk

Derivation of significant social network

Defining how nodes should be connected - i.e. the role of the edge set - was not immediately obvious. Previous work has employed a number of different topological definitions – both directed and undirected - for modelling discussions in online forums[1]. Because students sometimes tended to create new posts even when they meant to comment, we determined that a directed topology would not adequately capture the potential information flow and communication between learners.

Our guiding philosophy in formulating the network was to make the least assumptions about our data to determine which links depicting thread co-participation were "significant" (i.e., which ones were indicative of exchanges between learners that could indicate the existence, or potential for future existence, of underlying social relationships). Past research has used interaction time windows to determine which links to keep and which to remove in a network[2]. Because we did not have reliable a priori knowledge about what a reasonable time window would be, we instead turned to a significant network extraction model used to infer social networks in ecological settings[3] in order to inform our efforts.

The derivation of a significant social network proceeds as follows. For each sub-forum f, we first construct a bipartite graph mapping and represented by the *learner-to-thread* adjaceny matrix $B_f^{N_f \times T_f}$, where N_f represents the number of learners that explicitly posted and T_f the number of threads, both in sub-forum f. Each entry $b^{n,t} \in B_f^{N_f \times T_f}$ is an integer greater than or equal to 0, denoting the number of times learner $n$participated in thread t. Next, we compute a standard weighted one-mode projection of B_f to recover the *learner-to-learner* network, $L_f^{N_f \times N_f}$ [3]. Each entry $l^{i,j} \in L_f^{N_f \times N_f}$ is also an integer greater than or equal to 0, depicting the number of times that learner i or j co-participated in a particular thread t .

With a learner-to-learner network L_f, we are now tasked with identifying which edges in L_f (i.e., $l^{i,j} \in L_f$ s.t. $l^{i,j} \neq 0$) depict a *significant* interaction between two learners. Our goal is to generate a family of M sample networks against which we can test the significance of each $l_{i,j} \in L_f$. We start by noting that the observed learner-to-thread network B_f depicts each learner n's participation in a particular thread t. We can model this participation – i.e., each row n of B_f – as a draw from $Multinomial(k_n, p_n)$, where $k_n = \sum_{t \in T_f} b_{n,t}$ and $p_n = \left(p_{n,1}, \ldots., p_{n,T_f}\right)$ for $p_{n,t} = b_{n,t}/k_n$. It is important to note that $p = \{p_n\}_{i=1}^{N}$ represents the observed social relationships between learners as represented by the likelihood of each student's participation in a particular thread.

If we wish to test the significance of the observed edges – i.e., the observed social interactions – we must determine a mechanism for generating possible social networks that do not possess the same social patterns as the observed one. In order to explore alternative social structures, we first define a shuffling function σ such that each row of the sth sample learner-to-thread network B_f^s is drawn from $Multinomial\left(k_n, \sigma(p_n)\right)$ with k_n and p_n defined as above. We define σ such that it preserves learner n's proportional participation in different threads (e.g., the entropy of each p_n), but accounts for the possibility of participation in alternate threads. As an extension to Psorakis et al., we constrain σ to only shuffle each entry of p_n with a location (e.g. thread) t that has popularity greater than or equal to the least popular thread that learner n participated in, where thread popularity is defined as the number of posts it contains. This constraint is meant to reflect which threads learners could have possibly participated in, since in many cases, discussion threads only had a single or very small number of posts, and therefore, it is unrealistic to assume that a learner who participated primarily in popular threads may have also participated in isolated ones. Without this constraint, the shuffling allocates participation probabilities to a larger set of threads, increasing the likelihood – particularly for those individuals with low participation volumes but high proclivity to post in popular threads – that these one-off interactions are deemed "significant". Additionally, this constraint is informed by real-world discussions with participants from FOBS 1, some of who indicated that the popularity of a particular discussion thread often influenced their decisions to view or post. Therefore, the constrained shuffling more accurately captures learner behaviour and detects one-off participation in high-activity threads as insignificant (thereby, pruning more edges) when compared to its under-constrained counterpart.

With a sampling procedure in place, we generate each B_f^s and compute its one mode projection to arrive at the set of sampled learner-to-learner networks, i.e. $G = \left\{G_f^s\right\}_{s=1}^{M}$. We can then compare each entry $l_{i,j} \in L_f$ to $\frac{1}{M}\sum_s g_{i,j}$ for $g_{i,j} \in G_f^s$, computing the z-score and labeling as significant if the right-tailed p-value is less than 0.001 (we assumed a relatively small p-value threshold due to the sparsity of participation in the discussion threads). Our derived significant network is the collection of $l_{i,j} \in L_f$ labeled as significant by this procedure.

It is important to note that the determination of social significance by this procedure relies almost entirely on the frequency with which learners co-participate in threads. Frequency of forum participation has been investigated by others as a means of determining engagement and evaluating performance in educational settings[2]. Still, in a complex social setting such as an online course, the significance of communication is not entirely dependent on the frequency of co-participation, but also on the nature of the content exchanged. Developing automated ways of content-based significance testing is an opportunity for further research in MOOCs.

Supplementary figures

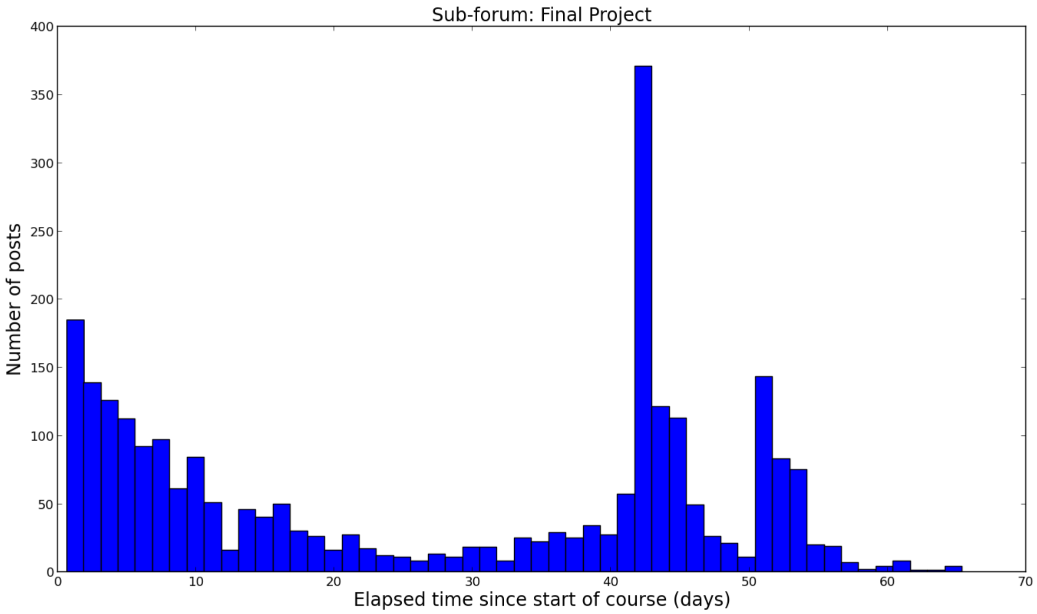

Figure S1: Forum post activity over time in the Final Project sub-forum of FOBS-1. The large peak around the end of week 6 corresponds to students posting last-minute questions about the final project submission deadline.

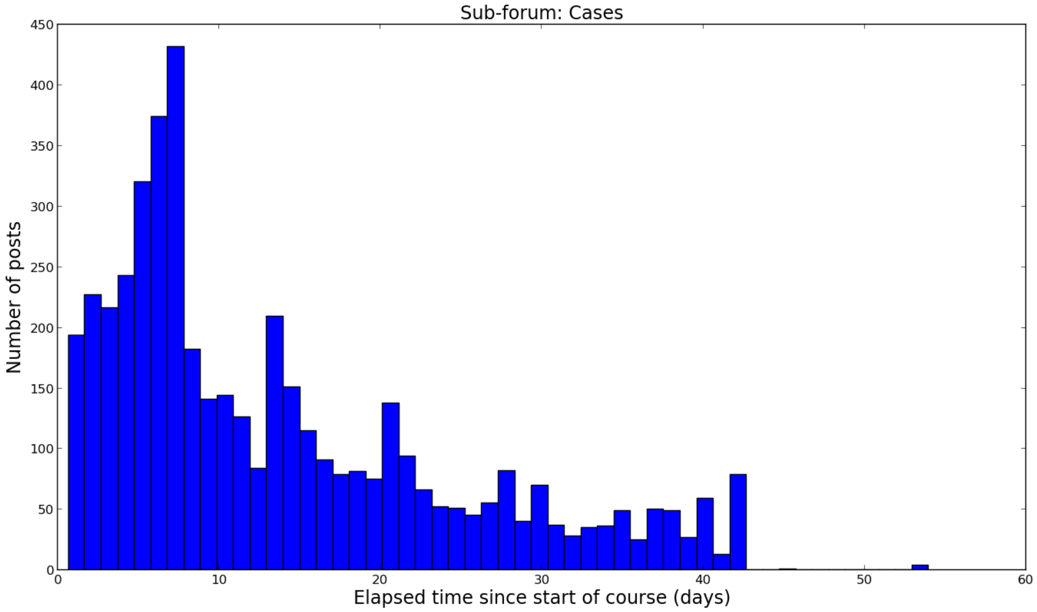

Figure S2: Forum post activity over time in the Cases sub-forum of FOBS-1. Like many of the other sub-forums, participation decreases as the course progresses, but there are still peaks of activity each week corresponding to the weekly case discussions.

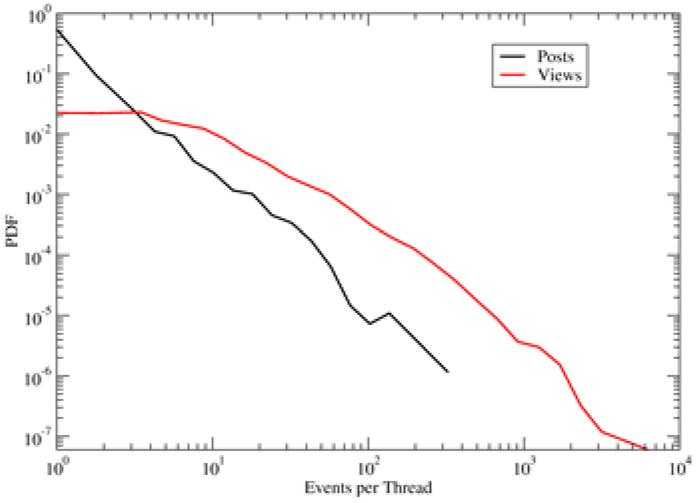

Figure S3: The number of views and posts per discussion threads across all sub-forums, in log-log scale, for FOBS-1. The charts suggest a fat-tailed distribution of views and posts across threads – i.e., the vast majority of discussion threads have very small numbers of posts and views, with a few threads harbouring high posting and viewing behaviour.

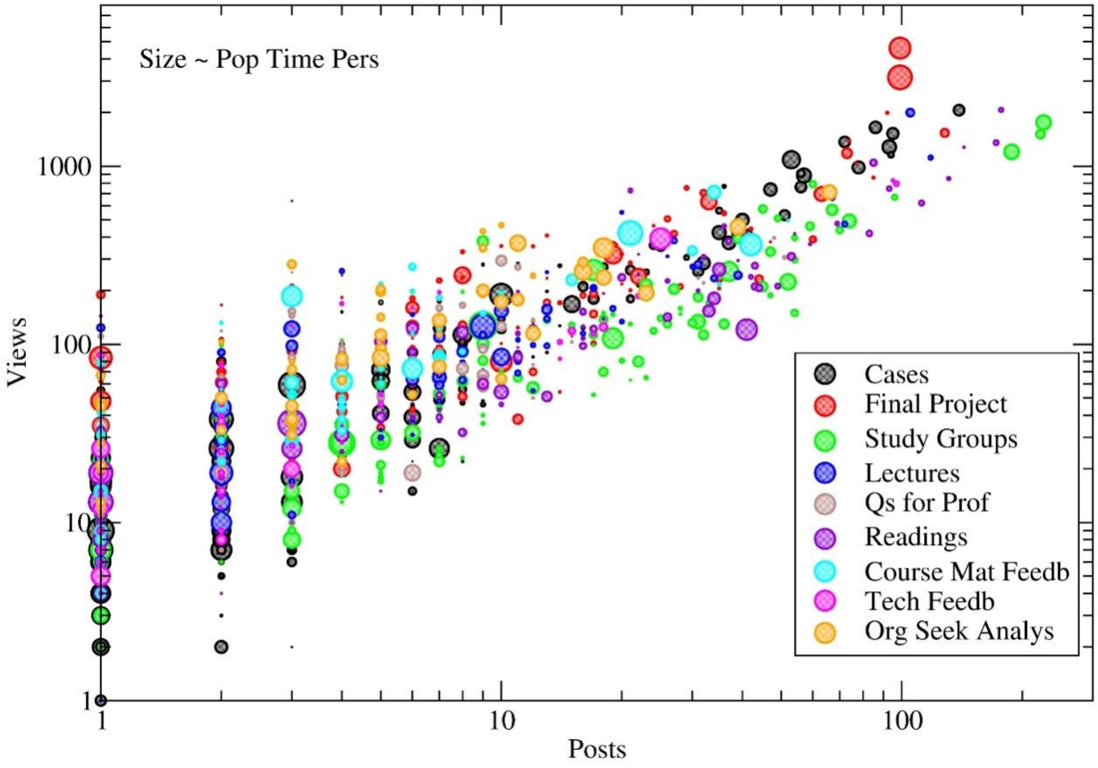

Figure S4: Comparison of posts and views for each thread in a particular sub-forum, denoted by the coloured circles shown here, for FOBS-1. The size of each circle indicates the “Popularity time persistence” of the corresponding thread, i.e., the amount of time that elapses before 90% of all posts are made to that thread (hence, small circles depict threads with very short lifespans).

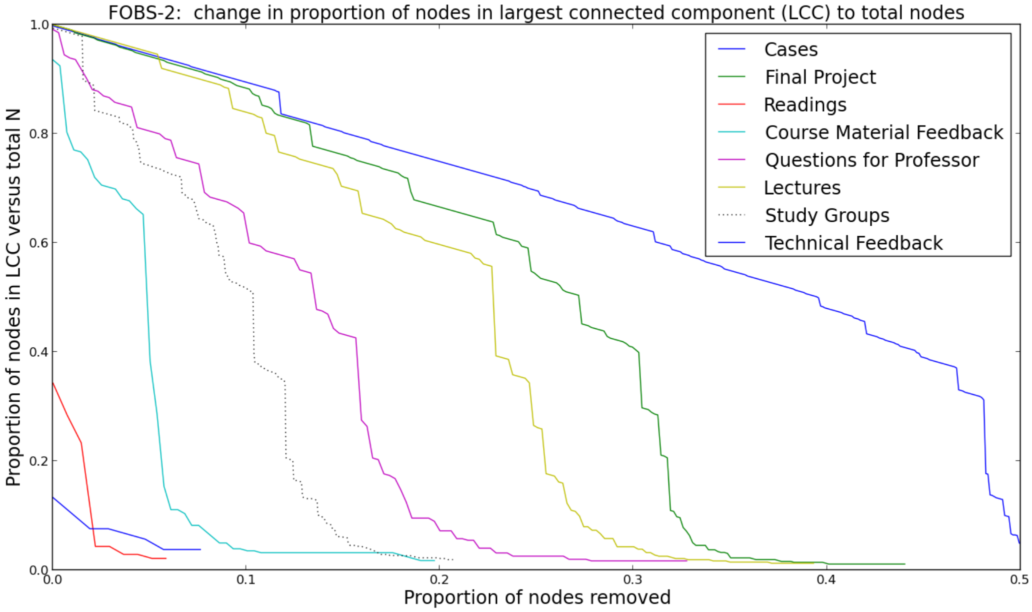

Figure S5: Communication vulnerability in the different sub-forums of FOBS-2. These trends are similar to those observed in FOBS-1.

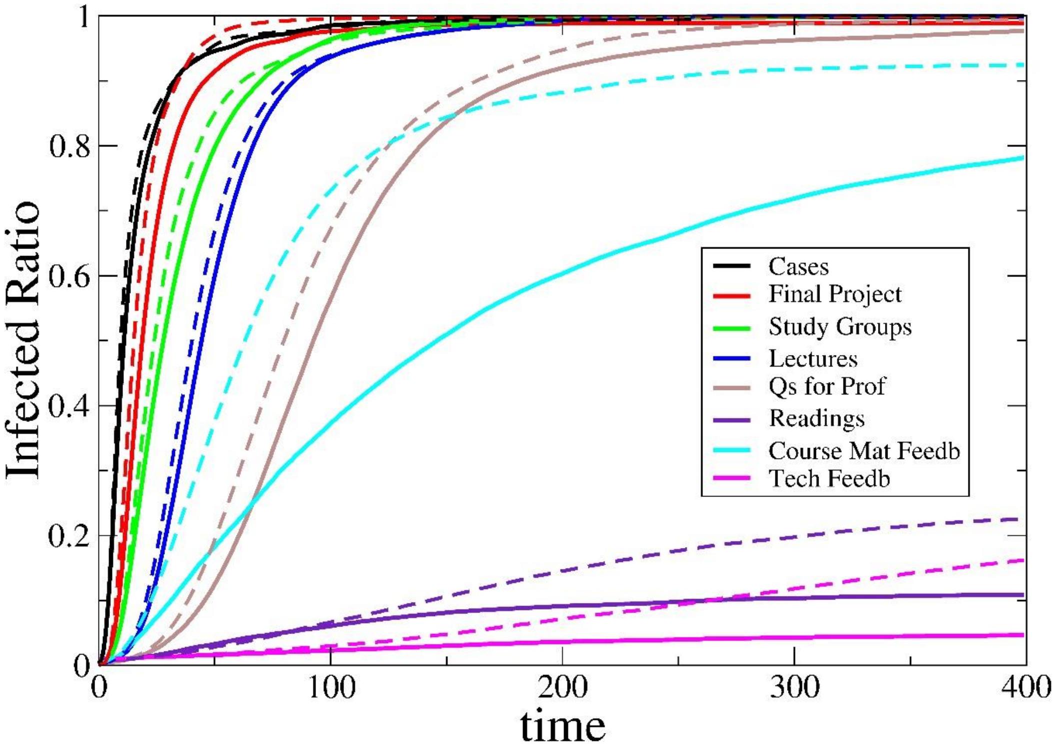

Figure S6: shows the percentage of infected nodes vs. simulation time for different networks in FOBS-2 (similar to those observed in FOBS-1). The solid lines show the results over the original network and the dashed lines for the degree-preserved shuffled network (configuration model).

Supplementary video

The video at the link below depicts the network vulnerability simulation for the Final Project sub-forum from FOBS-1. Each frame corresponds to one step in the algorithm, at which the node with the highest betweenness centrality is computed and disconnected from the graph.

https://www.dropbox.com/s/rvkd18dnuiyd02v/finalprojects.avi

References

1. Huerta-Quintanilla, R. . C.-L. E. &. V.-d. A. D., Modeling Social Network Topologies in Elementary Schools. *PLOS One* **8** (2), 1-9 (2013).
2. Holme, P. & Saramaki, J., Temporal networks. *Physics Reports* **519** (3), 97-125 (2012).
3. Psorakis, I., Roberts, S. J., Rezek, I. & Sheldon, B. C., Inferring social network structure in ecological systems from spatio-temporal data streams. *Journal of the Royal Society Interface* **9** (76), 3055-3066 (2012).
4. Davies, J. & Graff, M., Performance in e-learning: online participation and student grades. *British Journal of Educational Technology* **36** (4), 657-663 (2005).