Journal of Applied Engineering Mathematics

Volume 12, December 2020

Modelling censorship in complete graph social networks via the heat equation and probabilistic digital boundary conditions

Chayse Wright
Mechanical Engineering Depatrment
Brigham Young University
Provo, Utah 84602
chaysew@byu.edu

Abstract

In this paper, violent posts on online social networks are thermodynamically modelled via a digital adaption of the heat equation across a complete graph for a small N. This digital model then has a probabilistic, logarithmic, digital boundary condition applied at each user interface, allowing for safe opinion temperature dissipation, network user inclusion, and the prevention of echo chamber formation that is characteristic of an impermeable boundary condition.

Nomenclature

W	Considered time window of posts
i	Internal poster (Heat generation in the interior of the boundary)
j	External poster (Heat generation outside of the boundary)
T	Temperature
e	Graph edge
n	Node
K_n	Undirected, finite, complete graph of n nodes
P_{ji}	Probability of post being moderated
r	Random number for Bernoulli Trials

Introduction

Relationships between heat and social contagion are common parallels for researchers of social dynamics [1]. In this text, boundary conditions to prevent explosive and violent events among a population are studied by modelling opinion "temperature" among a network of people, where each person is connected to each other person, forming K_n, a complete graph of $\mathcal{N}^2$ connectivity. On most social media platforms, a given user of a social network makes a post, which is then characterized by a network moderation filter and assigned a score associated with how violent the post is. The platform then uses that filter to assign whether that post can be shared with others, completely preventing propagation of the violent opinion.

In the author's view, such a strict boundary condition leads to gradual opinion temperature buildup, eventually leading to violent action or a "meltdown". In this paper, the heat equation is adapted to a network of digital connections, and a probabilistic digital boundary condition is created that minimizes the risk of harmful opinion outbreak across a social network by retaining 'safe' user post sharing and minimizing the production of constant high temperature "echo chambers" that are siloed off to the rest of the public[2]. The boundary condition dictates the probability that a single post will be shared to other individuals based on a logarithmic scaling of the severity of the post and the temperature of the two users.

Thermodynamic Digital Social Networks

Continuous thermal approximations of social networks are ill suited to modelling digital interactions and therefore opinion temperature propagation that the networks entail. It is necessary to transform all aspects of the thermodynamic continuous model to a digital representation. Temperature for a user is modelled as a function of heat generation (i posting) and flux (j exposure). Temperature is measured from time t, the beginning of the window to $W + t$, the time at the end of the window, which will be held as the current time for the rest of this text.

$$T_i(W+t) = T_i(t) + \int_{\mathrm{t}}^{\mathrm{W+t}} \left[\mathrm{S}_i^{own}(s) + \sum_{j\neq i} F_{ji}(s) \right] ds, \quad (1)$$

where $\mathrm{S}_i^{\mathrm{own}}(s)$, is the intensity of the users own posts, given by the equation

$$S_i^{own}(s) = max('severity\ of\ posts\ in\ W\ at\ s') \times log(1 + 'post\ frequency\ in\ W\ at\ s') \quad (2)$$

where the severity is the model assigned violence severity score $S \in [0,1]$.

A rough intuition for the severity score can be developed as follows: “This is confusing” would receive a violence severity score of 0, whereas “I already know where you live and am currently travelling to your location to murder you” would receive a 1. Frequency is log normalized to account for the fact that low severity, high posting frequency action should not be as intense as high severity, low posting frequency activity. $F_{ji}(s)$ is the intensity of observed posts from all posters except themselves (i ≠ j). With sufficient individual posting frequency and a large network population we can modelled the intensity of observed posts as

$$F_{ji}(s) = P_{ji}(s) \times S_j(s), \quad (3)$$

where $P_{j\mathrm{i}}$ is the probability that user i will see the post of user j. $P_{j\mathrm{i}}$ in an unmoderated network is 1. For post moderation, we will impose an equation for P_{ij} that matches desired network sharing. S_j is the intensity of posts from user j.

Probabilistic Digital Boundary Condition

To establish the allowed flux between users and enable content moderation we establish a special boundary condition. This boundary condition must not only be effective at dissipating heat in high opinion temperature environments, but it should also do so in a manner that is not immediately discernible to users to discourage siloing and encourage user retention.

Thus, the addition of some probabilistic component allows for dissipation of heat in the lower temperature environment via discrete posts in a scaled fashion. This boundary must be effective in dealing with sudden high shock heat production events, which in the limit must lead to the emergency measure of effective siloing. Before further development of P_{ij} it is prudent to specify a boundary condition between nodes. Dirichlet conditions would lead to echo chambers, Neumann conditions allow for unchecked heat flow. The Robin boundary conditions describe heat convection, which is generally expressed as

$$\alpha T + \beta \frac{\partial T}{\partial n} = g,$$

where α and β are both parameters and g is some specified function. This general boundary condition can then be discretely represented on finite, undirected graphs as

$$\sum_{e \ni \mathrm{n}} \frac{dT_e}{d\mathrm{n}_e}(n) + \beta T(\mathrm{n}) = 0,$$

where T_e is a function on e and β is a parameter. Due to the connectivity of K_n, the derivative of T_e simplifies to $u(j) - u(i)$ leading to the following degeneration,

$$\sum_{\mathrm{j}\neq\mathrm{i}} \big(u(j) - u(i)\big) + \beta_{\mathrm{i}} u(i) = 0,$$

$$(n-1)\sum_{\mathrm{j}\neq\mathrm{i}} u(j) - (n-1)^2 u(i) + \beta_{\mathrm{i}} u(i) = 0. \quad (4)$$

For this to be utilized, it must apply to all n, and to solve for the nontrivial cases of βi ≠0, either the Robin term must be dropped, or the desirable aspects of the Robin BC grandfathered into a global parameter that exists in Kn. This global parameter should mimic the damping and spike-resistant traits that we desire from the Robin boundary condition. These characteristics will be presently developed in Eq. 5

$$P_{ji}(s) = \frac{1}{1 + exp\left(\delta\left(T_j(s) + T_i(s)\right) + \epsilon \cdot severity - \beta\frac{\partial T}{\partial n} + \gamma\log(1 + |F_{ji}|\right)}. \quad (5)$$

δ is a weighting parameter for sharing when both i and j have high T values. Set $\delta > 0$ to discourage sharing between high T posters. δ can be dynamically tuned according to moderator preference of echo chamber prevention.

ϵ is a weighting parameter that can be tuned to reduce the odds of severe posts being shared. The parameter β acts on the directional gradient, encouraging heat flow from hot to cold opinion temperature nodes. The parameter γ scales a logarithmic flux saturation term, logarithmically decreasing the probability of sharing posts with high intensity without hard cut off, providing an effective silo for heat spike events. All terms are included in x for the general sigmoidal $\frac{1}{1+e^x}$ form[3].

The general sigmoidal form has the advantage of never reaching zero, or 1, meaning that low severity posts only have a fifty percent score when $x = 0$, preventing total transmission of low severity posts across the entire network.

Results

To facilitate the modeling of a digital social network, the continuous thermodynamic equations were adapted into a discrete, stochastic agent-based simulation over K_{50}. Unlike continuous physical media where heat diffuses passively, this simulation required that "heat" (opinion intensity) be transmitted exclusively through discrete, time-staggered events representing social media posts. The network was initialized with a random history of low-severity interactions, establishing a baseline heterogeneous temperature distribution among users before the simulation window began ($t = 0$ to $t = 300$).

Post generation was modeled stochastically using a dynamic Beta distribution to determine the severity $S \in [0,1]$ of each content piece. The shape parameters of the distribution were coupled to the user's current normalized temperature. Users with low T produced a distribution with a heavy left tail, generating mostly benign content. As a user's temperature approached the maximum capacity $T_{max} = 5$, the tail was linearly shifted, skewing the probability density toward the right and increasing the likelihood of high-severity post generation. This mechanism effectively simulated the "self-heating" feedback loop where agitated users produce increasingly volatile content.

The propagation of this content across the network edges was done by a Bernoulli Trial, determined by the Probabilistic Digital Boundary Condition P_{ji}. For every potential connection in the K_{50} graph, a sharing probability was calculated based on the summation of penalty terms for $\epsilon, \delta,$ and γ.A pseudo-random number generator produced a value $r \in [0,1]$ for each edge; propagation occurred only if $r < P_{ji}$. As seen in Fig 1, this effectively acted as a stochastic filter, where high-severity posts between high-temperature users were statistically unlikely to propagate, while interactions flowing from high-temperature to low-temperature nodes were encouraged via the gradient term to encourage high temperature users to interact with moderate portions of the network.

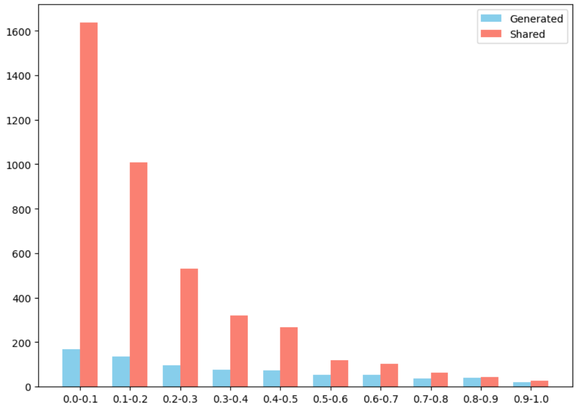


Figure 1: Severity S on x-axis, with the number of posts on the y-axis. Generated posts refers to posts coming from i and shared posts are those which succeed a Bernoulli Trial to be shared, coming from j .

To test the system's resilience, a severe "network shock" event was introduced at $t = 150$ where 80% of the nodes were instantaneously raised to the maximum temperature capacity[4]. The system's recovery was governed by a non-linear "overloaded heat sink" dissipation function ($D \propto \frac{T}{1+(T/K_{bend})^3}$), rather than linear Newtonian cooling. This bell-curve formulation meant that while dissipation was efficient at low temperatures, it dropped off precipitously as temperatures exceeded the inflection point K_{bend}. The results indicated that once the shock pushed users past this thermal threshold, the network entered a metastable high-temperature state, as the dissipation rate became insufficient to counteract even minimal self-generated heat flux. By user tuning to the network parameters, we can see in Fig 2 that the boundary condition proved resilient to critical events that generate a general and sudden increase in T across the entire network, effectively containing shock without total siloing of high T individuals.

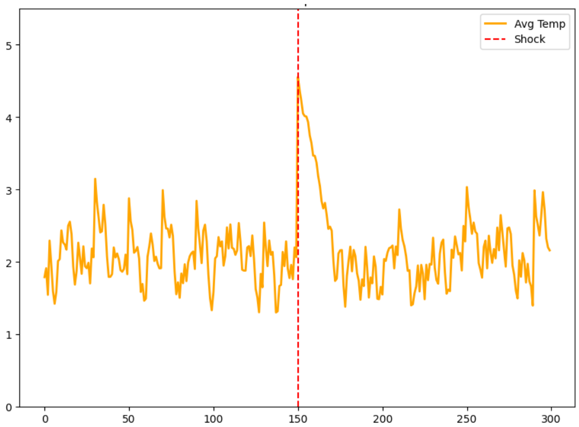


Figure 2: x-axis is the time in simulation t, the y-axis is the mean temperature T across all nodes in K_{50}. The network shock event can be seen at $t = 150$

Discussion

It is important for the moderator to supply a W value that captures a time window long enough to show real temperature dissipation and short enough to not include multiple temperature increasing events, creating long-term throttling of users for severe opinions that are no longer held. Note that no simple temperature gradient can be established across boundaries due to the nature of digital post interactions, necessitating the use of observed posts[5].

Including the β term in Pji is critical to this approach of moderation. By allowing high opinion temperature i to interact with low temperature j, moderated interactions ensue. This approach may increase dissatisfaction in generally low opinion temperature populations but serves the critical role of allowing high opinion temperature individuals to dissipate heat while minimizing the creation of high temperature groups, preventing violent event 'meltdown'

This model does not account for followers and assumes that all users have similar post reach. This could be corrected in future modelling by not using a complete graph but with localized networks with an uneven number and weighting of edges between users according to their reach. A limitation of this approach is that low severity posts are always modelled as having the maximum possible reach in Pji. This does not accurately represent modern observed interest feed generation algorithms. Perhaps the most important consideration for this paper is that many individuals that are exposed to high F_{ji} do not experience an increase in T and may exhibit negative flux on individuals with high amounts of Si own.

Conclusions

Discretization of the heat equation into a stochastic, agent-based model on a complete graph successfully demonstrated the viability of thermodynamic analogies for regulating digital social networks. The implementation of the Probabilistic Digital Boundary Condition confirmed that severity-weighted, stochastic filtering can effectively manage baseline opinion temperatures without resorting to absolute censorship, thereby preserving network connectivity while dampening the formation of low-temperature echo chambers. However, improper tuning of model parameters in the non-linear dissipation response creates a metastable high-temperature state following synchronized network shocks. Further investigations will necessitate consideration of user reach. An additional area of investigation that is necessary for comprehensive modelling is the inclusion of incomplete, sparse, or directed graphs in the simulation.

Acknowledgements

BYU Department of Physics and Astronomy, Dr. Eric Hirschmann for convincing me that trying to use boundary conditions and phase changes with general relativity was a nontrivial and unproven endeavour.

AI Disclaimer: Google Gemini 3 Pro, Grok 4 for rendering code and visualization formatting.

References

[1] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Rev. Mod. Phys.*, vol. 81, no. 2, pp. 591–646, May 2009

[2] R. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete structures," in *Proc. 19th Int. Conf. Mach. Learn.*, 2002, pp. 315–322.

[3] R. Hegselmann and U. Krause, "Opinion dynamics and bounded confidence: Models, analysis, and simulation," *J. Artif. Soc. Simul*., vol. 5, no. 3, 2002.

[4] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," Proc. Nat. Acad. Sci. USA, vol. 111, no. 24, pp. 8788–8790, Jun. 2014.

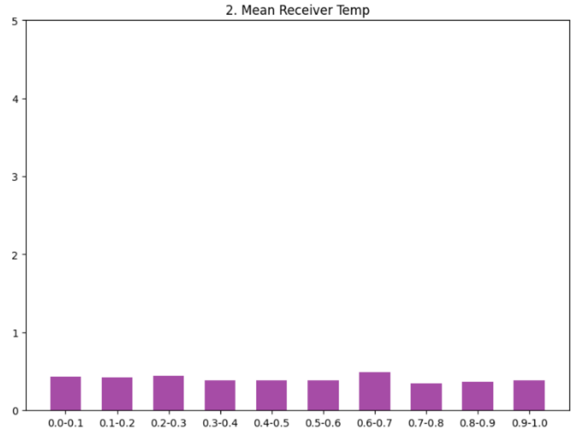
[5] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," Proc. Nat. Acad. Sci. USA, vol. 99, no. suppl 3, pp. 7280–7287, May 2002.

Appendix

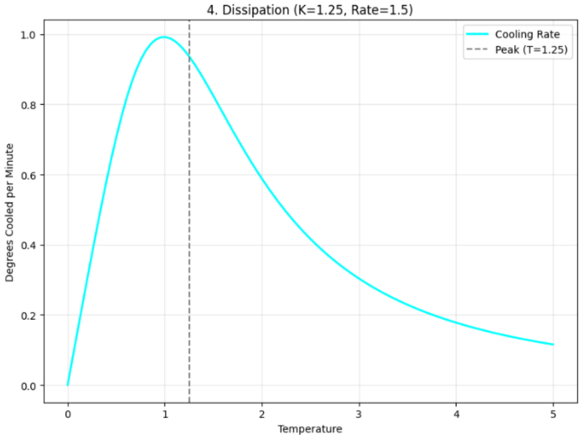
The full animation for the simulation discussed in this paper along with additional details can be accessed at:
https://youtu.be/7Pxxj5InVQM

All code for simulation and rendering can be accessed at:
https://colab.research.google.com/drive/1wEq1qe_x5FCmO1mrwU2KQSeMc0eagkqN?usp=sharing

1. Extra simulation plots showing mean receiver temp of i based on j post severity.



2. Dissipation curve.


4. Dissipation (K=1.25, Rate=1.5)
Cooling Rate
Peak (T=1.25)
Degrees Cooled per Minute
Temperature