A Social Balance Approach to Multi-Agent Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) can be deployed in situations where they process positive/negative interactions with other agents. We study how this is done under the sociological framework of social balance, which explains the emergence of one faction or multiple antagonistic factions among agents. Across different LLM models, we find that balance depends on the (i) type of interaction, (ii) update mechanism, and (iii) population size. Across (i)-(iii), we characterize the frequency at which social balance is achieved, the justifications for the social dynamics, and the diversity and stability of interactions. Finally, we explain how our findings inform the deployment of agentic systems.

1 Introduction

Large Language Models (LLMs) have shown language understanding (Xi et al., 2023; Wei et al., 2022a,b) and reasoning (Yao et al., 2023; Bang et al., 2023) abilities. In multi-agent settings, they can express complex opinion dynamics (Chuang et al., 2024a,b; Cisneros-Velarde, 2025). These abilities are influenced by biases (Liang et al., 2023) from pre-training and/or alignment. These biases also influence how LLMs display sociopsychological qualities (Zhou et al., 2024), express (Serapio-Garcia et al., 2023) or assess (Rao et al., 2023) personality traits and values or judgment (Wang et al., 2024; Sorensen et al., 2024).

Given the abilities and biased behavior of LLMs, it is relevant to understand how they assess positive and negative interactions for applications where their responses or actions depend on such assessments (Guo et al., 2024). For example, LLMs can participate in social media (Leskovec et al., 2010), where establishing negative ties have great influence on opinion mobilization (Kumar et al., 2018); negative ties are also associated to information diffusion and centrality of individuals (Ghasemian and Christakis, 2024). Section 5 further explains the application to social media. Thus, contributing to the literature on multi-agent LLMs in sociologically relevant scenarios (Gao et al., 2024), we study the question: *how do LLM agents understand and update positive and negative interactions among themselves?* To the best of our knowledge, this question is underexplored.

To study this problem concretely, we consider the well-established sociological framework of *social balance*. Social balance dictates the positive and negative interactions that must be held within *every* three individuals or *triad* in a population (e.g., a social network) for the emergence of *factions*. In particular, *structural balance* (Heider, 1946) characterizes the *signed* interactions that lead to the emergence of one *faction* where all individuals have positive interactions with each other, or two factions where interactions are only negative across factions. Within a triad, structural balance is dictated by so-called "Heider's rules": (i) the enemy of an enemy is a friend; (ii) the friend of an enemy is an enemy; (iii) the enemy of a friend is an enemy; (iv) the friend of a friend is a friend. It is argued that when one of these rules is violated, it generates a *cognitive dissonance* which the individual must resolve (Festinger, 1957; Ric, 2015). Another notion of social balance is *clustering balance* (Davis, 1967), which occurs when rule (i) is relaxed within a triad of agents, i.e., three agents can be enemies of each other. This allows the emergence of more than two antagonistic factions. Fig. 1 illustrates all balanced triads.

Thus, in this paper, our main focus is to study the appearance of social balance in a group of *three* LLM agents continuously updating their interactions with each other. It is logical to first establish results for this fundamental case because a triad *is* the basis of Heider's rules. After this, we also study populations of multiple triads (i.e., with more than three agents).

We characterize the interactions along two dimensions. The first is the *type of interaction*: friendly and unfriendly ties among agents can be expressed as positive and negative *relationships* with each other, or *appraisals* or *opinions* of each other. These types are present in both theoretical and empirical sociological literature; e.g., (Szell et al., 2010; Leskovec et al., 2010; Cisneros-Velarde and Bullo, 2020). The second dimension is the *update mechanism* for the agents' interactions. Motivated by mathematical sociology (Mei et al., 2019), we consider *homophily* and *influence*. Assume Agent i wants to update its tie to Agent j. Homophily is when Agent i does a comparison between (a) how it and Agent j relate to the other agents, and (b) how it directly relates to Agent j. Influence is when Agent i considers, instead of (b), how Agent j directly relates to Agent i.

Our **main contribution** is to show that diverse LLM models—of different size and family, having alignment or not, being a reasoning model or not—are able to achieve social balance after repeated interactions: showing the emergence of one faction or multiple antagonistic factions in a social group.

For a population of a single triad:

The *type* of social balance (structural or clustering), *diversity* of *sign configurations* of the achieved balanced triads (Fig. 1), and *frequency* of achieving social balance differ across LLM models and types of interactions and update mechanisms. Generally, larger models achieve social balance more frequently and in its strongest (structural) type. Models with strong alignment have a preference for all-positive or all-negative interactions.

LLM models process update mechanisms differently: the *same* update mechanism can have different and even opposite effects on the number of positive/negative interactions and on the diversity of balanced triads.

Remarkably, a model that achieves social balance more often does not necessarily have more diversity of socially balanced triads or more *stable* (i.e., less changing) interactions.

Finally, LLMs show knowledge of sociopsychological processes that lead to social balance: the keyword "dissonance" is used by agents to justify new interactions that *do not contradict* Heider's rules—thus agreeing with sociological theory. A reasoning model approaches the update of its interactions more mathematically. □

For a population of multiple triads:

Unlike the single triad case, smaller models achieve social balance more frequently across most settings. Thus, achieving social balance is not a property that *absolutely* improves with model size. Moreover, more models achieve clustering balance. All of this is possible evidence that social interactions are updated under *different* dynamics in larger populations. Finally, the keyword "dissonance" is largely absent across LLM responses. □

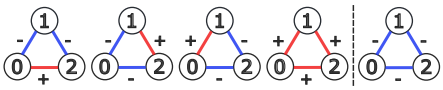

Figure 1: **Sign configurations of a balanced triad.** Consider a triad as a graph with three nodes corresponding to the agents 0, 1, and 2. The first four triads define structural balance (e.g., in the first triad: agent 2 is an enemy of 1, who in turn is an enemy of 0; therefore, 2 is a friend of 0). All five triads define clustering balance.

We further discuss our work's motivation in App. A. The related literature is in App. B.

2 Problem Setting

We consider a population of LLM agents of the same model, type of interaction (relationship, appraisal, or opinion), and update mechanism (homophily or influence). Initially at $t = 0$, every agent has positive or negative interactions. Interactions are directional and need not be symmetrical (e.g., Agent i has a positive opinion of j, while j has a negative opinion of i). For a single triad population, we initialize the interactions according to all 2^6 possible sign distributions. We run ten simulations per sign distribution. For a population of multiple triads (size larger than three), we do ten random initialization for the interaction signs. At every iteration $t = 1, \ldots, 10$, regardless of the population size, all agents simultaneously update their interactions toward every other agent; see Fig. 2. All agents are asked to justify the update of their interactions. A population achieves social balance when every triad in it satisfies Fig. 1 (all interactions are symmetrical) at iteration $t = 10$. Further mathematical, experimental, and prompt details are in App. G.

3 Analysis: Single Triad Population

We first study a population of three LLM agents. Supplementary information is in App. C.

3.1 Achieving social balance

We find that all LLM models achieve the well-established sociological phenomenon of social bal-

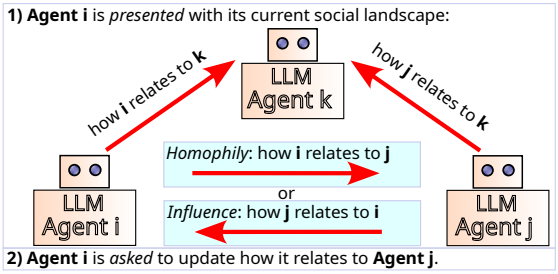

Figure 2: **Agent i updates its interaction toward Agent j.** The update depends on how both relate to every other Agent k in the population.

	Rel.		Apprais.		Opin.		
	H	I	H	I	H	I	type
gpt-oss-120b	66.88	36.72	96.88	26.41	95.31	23.75	Both
Llama 3 70B	**76.56**	**87.50**	**100.00**	**54.69**	18.75	26.88	S
Llama 3 8B	5.63	25.62	44.22	0.00	**98.44**	**56.25**	Both
Mistral 7B	—	9.38	23.44	7.81	10.94	7.81	C
Gemma 3 4B	52.81	34.38	43.75	27.50	27.66	15.63	C

Table 1: **Frequency (percentage %) of simulations that result in social balance.** H: homophily, I: influence. We indicate whether each model has *every* setting resulting *exclusively* in structural (S)/clustering (C) balance, or if both types appear. Only Mistral 7B refuses to provide an answer under relationships with homophily.

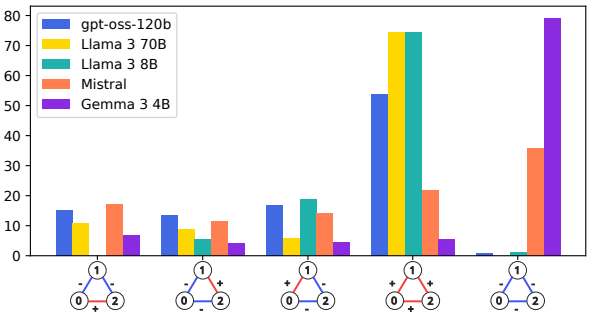

Figure 3: **Frequency (%) of balanced triads.** For each model, we present the histogram of types of balanced triads that emerge at the end of all simulations.

ance; see Tables 1 and 2, and Fig. 3. Moreover, the *type* of social balance, the *frequency* at which it is achieved, and the *diversity* of achieved balanced triads differ among models. Thus, *social balance can be used as an appropriate tool to characterize the social dynamics embedded in LLMs.*

Regarding the **type** of social balance, gpt-oss-120b and Llama 3 70B achieve structural balance (no triad has all negative interactions) the most across all types of interactions and update mechanisms, whereas the smaller Mistral 7B and Gemma 3 4B favor clustering balance (Tables 1 and 2, and Fig. 3). Moreover, gpt-oss-120b and the Llama 3 family lead to an overwhelming presence

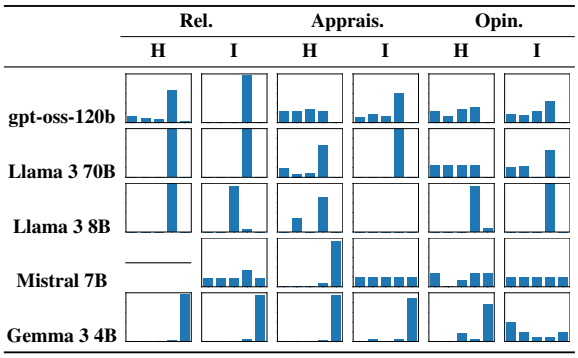

Table 2: **Diversity of balanced triads.** Each histogram has the x-axis corresponding to the five balanced triads as in Fig. 3 and the y-axis going from 0% to 100% frequency. The setting is similar to Table 1.

of all-positive triads (i.e., one faction), whereas Gemma 3 4B to all-negative triads (i.e., three factions) (Fig. 3). Mistral 7B is more evenly distributed. Thus: *(i) largest models are likely to enforce the strongest concept of social balance, and (ii) models with strong alignment have a preference for extreme positive/negative social outcome.* [1]

Regarding the **frequency** of achieving social balance, gpt-oss-120b and the Llama 3 family achieve social balance in the majority of simulations for most settings (boldface/underline in Table 1), whereas this is not true for other models (Table 1). We also find two surprising observations relating the frequency and **diversity** of achieved balanced triads. First, in the case of opinion interactions, Llama 3 8B achieves social balance more often than gpt-oss-120b and Llama 3 70B (Table 1); however, these two models achieve more diverse sign configurations (Table 2). Similarly, Mistral 7B has more diversity in settings where it achieves social balance less often. Surprisingly, gpt-oss-120b achieves more diversity and frequency together under appraisals than other models. Thus: *(i) achieving social balance more often does not imply achieving it with more diversity of interactions and vice versa; (ii) model size is not a reliable predictor for achieving more diverse social outcomes.*

One last observation about **diversity** of balanced triads: for a *fixed* type of interaction (Table 2), *the same update mechanism can (i) affect the diversity of balanced triads differently across models, and (ii) incite more positive interactions in one model and more negative ones in another.*

[1] To the best of our knowledge, only Mistral 7B has not undergone strong safety alignment.

	Rel.		Apprais.		Opin.	
	H	I	H	I	H	I
gpt-oss-120b-J	0.03	×	0.71	0.07	0.76	0.01
gpt-oss-120b-R	0.66	0.15	5.14	0.55	2.96	0.12
Llama 3 70B	×	×	13.80	5.28	6.28	1.16
Llama 3 8B	×	×	4.10	×	1.90	0.01
Gemma 3 4B	×	×	4.60	0.23	3.06	0.94

Table 3: **Frequency (percentage %) of responses with the keyword "dissonance".** gpt-oss-120b: response justification (J) and reasoning (R). "×" indicates 0.00.

3.2 Stability of interactions

Having shown that social dynamics of LLM agents can *lead* to social balance, we study its *stability*.

Our *first* stability criterion analyzes how often an initially balanced triad does not change any of its interactions after ten iterations. The stability (decreasing) order is: Mistral 7B (stable 84.00% of cases) $\succ$ Gemma 3 4B $\succ$ gpt-oss-120b $\succ$ Llama 3 70B $\succ$ Llama 3 8B (16.67%). Our *second* stability criterion measures how often LLM agents do not change the sign of their interactions in the last five iterations, regardless of the final triad being balanced or not. The stability (decreasing) order is: Mistral 7B (stable 99.78% of cases) $\succ$ Llama 3 70B $\succ$ Gemma 3 4B $\succ$ gpt-oss-120b $\succ$ Llama 3 8B (49.95%).

Our stability results bring a new perspective on social dynamics: *(i) the model that achieves social balance more often is not necessarily the one with more stable interaction updates*; and *(ii) LLMs are able to establish substantially stable social interactions despite their intrinsic stochastic nature.* Fig. 4 in App. C shows stability examples.

3.3 The justification of interaction updates

For the sake of interpretability, we focus on finding evidence of LLM agents *expressing* an intention to achieve social balance. Particularly, we look for relevant keywords in every justification given by every agent when updating its social interactions. In this section we only consider the keyword "dissonance" which is argued to occur when Heider's rules are violated (Section 1); other keywords are analyzed in App. C. The keyword "dissonance" is found in the Llama 3 family and Gemma 3 4B under appraisals and opinions; see Table 3. Coincidentally, appraisals and opinions lead to a larger diversity of balanced triads than relationships (Table 2). Moreover, this keyword appears most often in Llama 3 70B in the same setting that leads to social balance with 100% frequency (Table 1). Remarkably, "dissonance" is used to justify interaction updates that *explicitly resolve* a violation of Heider's rules (enforcing social balance), though in Gemma 3 4B the opposite is also often observed; see App. I. Interestingly, gpt-oss-120b uses a mathematical approach in order to enforce Heider's rules—probably due to being a reasoning model (App. I).

Thus, *LLMs can show knowledge of sociopsychological processes that lead to social balance, and this could benefit both diversity and frequency of balanced triads.*

4 Analysis: Population of Multiple Triads

All LLM models achieve social balance for populations of 20 and 120 triads, i.e., of six and ten agents, respectively; see Table 5 in App. D. Social balance is achieved if *every* single triad is balanced.

Two notable differences exist compared to the single triad case. First, the smallest models achieve social balance more often than the largest ones across most settings, even when the larger ones do not (Table 5). Second, only gpt-oss-120b and Llama 3 70B achieve all-positive interactions, and only in one setting. Llama 3 70B is now able to achieve clustering balance (Table 5).

Thus, our results indicate that, *in the more complex social scenario of populations with multiple triads, the LLM models display social dynamics different from the single triad case when updating interactions.* Moreover, *unlike the single triad case, a larger model does not necessarily achieve social balance more often and in a stronger type.*

Llama 3 70B and gpt-oss-120b only report "dissonance", and in a tiny amount: no more than 0.17% per setting in Llama 3 70B—compared to the 13.80% for a single triad (Table 3)—and considerably less in gpt-oss-120b; see Table 6 in App. D. *This supports the existence of different social dynamics when LLMs belong to larger populations.* Other keywords are analyzed in App. D.

5 Further Discussion

We provide a detailed example of how our results inform an agentic social media application in App. E. Our choice of LLM models, update of interaction signs, display of social information, use of LLMs without impersonation, and number of interaction rounds are further discussed in App. F. A discussion on our prompts is in App. H. Finally, our work's further motivation is discussed in App. A.

Limitations

All experiments have the temperature hyperparameter set to zero, which is a legitimate setting for applications where consistency on LLMs' outputs are desired; nonetheless, achieving social balance could be sensitive to this hyperparameter. Our experiments in the multiple triads population case are time consuming, which is why we performed less simulations compared to the population of a single triad. Our experiments consider a diverse group of open-source LLMS, nonetheless, they could also benefit from closed-source LLMs.

Ethical Considerations

The strong tendency of LLM models to achieve social balance could have a negative impact on the deployment of LLMs as bots in social media—this is further explained in App. E.

References

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org.

Omid Askarisichani, Ambuj K Singh, Francesco Bullo, and Noah E Friedkin. 2020. The 1995-2018 global evolution of the network of amicable and hostile relations among nation-states. *Communications Physics*, 3(1):215.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

David Camacho, Ángel Panizo-LLedot, Gema Bello-Orgaz, Antonio Gonzalez-Pardo, and Erik Cambria. 2020. The four dimensions of social network analysis: An overview of research methods, applications, and software tools. *Information Fusion*, 63:88–120.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert D. Hawkins, Sijia Yang, Dhavan V. Shah, Junjie Hu, and Timothy T. Rogers. 2024a. Simulating opinion dynamics with networks of LLM-based agents. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Yun-Shiuan Chuang, Siddharth Suresh, Nikunj Harlalka, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. 2024b. The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents. *Preprint*, arXiv:2311.09665.

Pedro Cisneros-Velarde. 2025. Biases in opinion dynamics in multi-agent systems of large language models: A case study on funding allocation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1889–1916, Albuquerque, New Mexico. Association for Computational Linguistics.

Pedro Cisneros-Velarde and Francesco Bullo. 2020. Signed network formation games and clustering balance. *Dynamic Games and Applications*, 10:783–797.

Pedro Cisneros-Velarde, Kevin S. Chan, and Francesco Bullo. 2021a. Polarization and fluctuations in signed social networks. *IEEE Transactions on Automatic Control*, 66(8):3789–3793.

Pedro Arturo Cisneros-Velarde, Noah E. Friedkin, Anton V. Proskurnikov, and Francesco Bullo. 2021b. Structural balance via gradient flows over signed graphs. *IEEE Transactions on Automatic Control*, 66(7):3169–3183.

James A. Davis. 1967. Clustering and structural balance in graphs. *Human Relations*, 20(2):181–187.

Patrick Doreian and Andrej Mrvar. 2015. Structural balance and signed international relations journal of social structure. *Journal of Social Structure*, 16(01).

David Easley and Jon Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

Shangbin Feng, Herun Wan, Ningnan Wang, Zhaoxuan Tan, Minnan Luo, and Yulia Tsvetkov. 2024. What does the bot say? opportunities and risks of large language models in social media bot detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3580–3601, Bangkok, Thailand. Association for Computational Linguistics.

Leon Festinger. 1957. A theory of cognitive dissonance. Stanford University Press.

Janet S. Fink, Heidi M. Parker, Martin Brett, and Julie Higgins. 2009. Off-field behavior of athletes and team identification: Using social identity theory and balance theory to explain fan reactions. *Journal of Sport Management*, 23(2):142 – 155.

Noah E. Friedkin. 1998. *A Structural Theory of Social Influence*. Cambridge University Press.

Anna Gallo, Diego Garlaschelli, Renaud Lambiotte, Fabio Saracco, and Tiziano Squartini. 2024. Testing structural balance theories in heterogeneous signed networks. *Communications Physics*, 7(1):154.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1259.

Amir Ghasemian and Nicholas A. Christakis. 2024. The structure and function of antagonistic ties in village social networks. *Proceedings of the National Academy of Sciences*, 121(26):e2401257121.

R. Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. 2004. Propagation of trust and distrust. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, page 403–412, New York, NY, USA. Association for Computing Machinery.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Frank Harary. 1953. On the notion of balance of a signed graph. *Michigan Mathematical Journal*, 2(2):143 – 146.

Ahmed Hassan, Amjad Abu-Jbara, and Dragomir Radev. 2012. Extracting signed social networks from text. In *Workshop Proceedings of TextGraphs-7: Graph-based Methods for Natural Language Processing*, pages 6–14, Jeju, Republic of Korea. Association for Computational Linguistics.

Fritz Heider. 1946. Attitudes and cognitive organization. *The Journal of Psychology*, 21(1):107–112.

Junjie Huang, Huawei Shen, Liang Hou, and Xueqi Cheng. 2021. Sdgnn: Learning node representation for signed directed networks. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI-21, page 555–564. Association for the Advancement of Artificial Intelligence.

Zexi Huang, Arlei Silva, and Ambuj Singh. 2022. Pole: Polarized embedding for signed networks. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 390–400, New York, NY, USA. Association for Computing Machinery.

Amiyaal Ilany, Adi Barocas, Lee Koren, Michael Kam, and Eli Geffen. 2013. Structural balance in the social networks of a wild mammal. *Animal Behaviour*, 85(6):1397–1405.

Krzysztof Kulakowski, Przemyslaw Gawronski, and Piotr Gronek. 2005. The heider balance: a continuus approach. *International Journal of Modern Physics C*, 16(05):707–716.

Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference*, page 933–943. International World Wide Web Conferences Steering Committee.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 1361–1370, New York, NY, USA. Association for Computing Machinery.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. EconAgent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15523–15536. Association for Computational Linguistics.

Percy Liang and 1 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Zeev Maoz, Lesley G. Terris, Ranan D. Kuperman, and Ilan Talmud. 2007. What is the enemy of my enemy? causes and consequences of imbalanced international relations, 1816–2001. *The Journal of Politics*, 69(1):100–115.

Seth A. Marvel, Jon Kleinberg, Robert D. Kleinberg, and Steven H. Strogatz. 2011. Continuous-time model of structural balance. *Proceedings of the National Academy of Sciences*, 108(5):1771–1776.

Wenjun Mei, Pedro Cisneros-Velarde, Ge Chen, Noah E. Friedkin, and Francesco Bullo. 2019. Dynamic social balance and convergent appraisals via homophily and influence mechanisms. *Automatica*, 110:108580.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Michael Moore. 1978. An international application of heider's balance theory. *European Journal of Social Psychology*, 8(3):401–405.

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatGPT assess human personalities? a general evaluation framework. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Francois Ric. 2015. Social cognition. In James D. Wright, editor, *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, second edition edition, pages 204–209. Elsevier, Oxford.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Greg Serapio-Garcia, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

Guodong Shi, Claudio Altafini, and John S. Baras. 2019. Dynamics over signed networks. *SIAM Review*, 61(2):229–257.

Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947.

Michael Szell, Renaud Lambiotte, and Stefan Thurner. 2010. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641.

Jiliang Tang, Yi Chang, Charu Aggarwal, and Huan Liu. 2016. A survey of signed network mining in social media. *ACM Comput. Surv.*, 49(3).

Vincent Antonio Traag, Paul Van Dooren, and Patrick De Leenheer. 2013. Dynamical models explaining social balance and evolution of cooperation. *PLOS ONE*, 8(4):1–7.

Arnout Van de Rijt. 2011. The micro-macro link for the theory of structural balance. *The Journal of Mathematical Sociology*, 35(1-3):94–113.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Arch G. Woodside. 2004. Advancing means–end chains by incorporating heider's balance theory and fournier's consumer–brand relationship typology. *Psychology & Marketing*, 21(4):279–294.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, and 10 others. 2023. The rise and potential of large language model based agents: A survey. *Preprint*, arXiv:2309.07864.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shuang-Hong Yang, Alexander J. Smola, Bo Long, Hongyuan Zha, and Yi Chang. 2012. Friend or frenemy? predicting signed ties in social networks. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, page 555–564, New York, NY, USA. Association for Computing Machinery.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2024. CompeteAI: Understanding the competition dynamics of large language model-based agents. In *Forty-first International Conference on Machine Learning*.

Xiaolong Zheng, Daniel Zeng, and Fei-Yue Wang. 2015. Social balance in signed networks. *Information Systems Frontiers*, 17(5):1077–1095.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

A Further Motivation for Our Work

The focus on social balance. Our paper aims to provide an answer to the *general question* of understanding how LLMs process signed interactions in a multi-agent setting. The fact that social balance *emerges* from the evolution of interactions is remarkable—when running our experiments there was *no prior guarantee or knowledge* that social balance would emerge. Therefore, given that (i) social balance is a well-established concept in the sociological literature and that (ii) it actually emerges in multi-agent LLM systems, we decided to place it as the *scope* of our paper.

Even within this scope, our work shows that LLM models display a variety of behaviors on *how* they achieve social balance (Sections 3 and 4), thus showing that social balance is a framework *rich enough* to capture diverse social dynamics embedded in LLMs.

We also remark that previous works have studied LLMs in social scenarios such as opinion dynamics in multi-agent systems (App. B). Our work *expands* this literature to the dynamics of sociologically relevant *signed* interactions across LLMs.

A connection with size complexity. It is known that some LLMs' abilities can improve and new ones can emerge as LLMs grow in size (Wei et al., 2022b). Our contributions show this is not necessarily true for collective emergent properties such as achieving social balance.

A connection with alignment. Our work contributes to the characterization of alignment in LLMs (Liang et al., 2023): *how* an LLM model achieves social balance could be used as a metric of *how* alignment affects social dynamics. For example, our results show that an LLM model that underwent a strong alignment process has either an extremely positive or negative outlook in their interactions—this behavior is not found in the only tested model without strong alignment (Section 3). Note that in our work LLMs are only affected by "built-in" biases, i.e., the default ones coming from alignment and/or training data, since there is no assigned persona (Aher et al., 2023) that could introduce additional biases (Salewski et al., 2023; Chuang et al., 2024a).

Our focus on LLM-to-LLM interactions. Our work only considers LLM-to-LLM interactions because it is a **controlled environment to analyze LLM multi-agent behavior**. This setting is also relevant for two reasons:

1. We believe it will be **more common** to find LLM-to-LLM interactions in social media in the near future, e.g., more users may have LLM agents posting and extracting information for them.

2. In our work, **the LLM does not know whether it is interacting with a human or another LLM**: nowhere in the prompts given to the LLMs it is indicated the nature of the "individuals" the LLM is interacting with (Apps. G.5 and G.6). In other words, the LLM is *agnostic* to the identity of the agents that interact with it. Thus, when an LLM explicitly changes its interactions in order to satisfy Heider's rules (e.g., by reducing cognitive dissonances; Section 3.3), **the same behavior is expected to occur even if humans were involved as interacting agents**.

B Related Literature

Dynamics of LLM interactions in multi-agent settings. To the best of our knowledge, no prior work has focused on how LLMs update different types of *signed* (positive or negative) interactions (i) without taking a specific persona, (ii) by only observing the current state of such interactions, and (iii) under the framework of social balance. The works (Chuang et al., 2024a,b) focus on how opinions spread and change among LLMs role-playing different persona, and the effect caused by injecting external biases and by the degree of impersonation. The work (Zhou et al., 2024) studies the display of social intelligence in role-playing LLMs, and, among different variables, explicitly considers the effect of family, friend, romantic, acquaintance, and stranger relationships among the agents. The work (Cisneros-Velarde, 2025) takes the LLM as is, and studies the conflict among biases present during opinion exchange. Such work considers opinions that display positive and negative connotations.

Dynamic modeling of signed interactions. Several works in mathematical sociology study how agents process signed interactions. One line of research studies how established positive and negative ties among agents affect the diffusion of opinions; e.g., social balance is known to lead to polarization of opinions (Shi et al., 2019; Cisneros-

Velarde et al., 2021a). Another line of research studies models that evolve signed interactions among agents toward structural balance (Kulakowski et al., 2005; Marvel et al., 2011; Traag et al., 2013; Mei et al., 2019; Cisneros-Velarde et al., 2021b) and clustering balance (Van de Rijt, 2011; Cisneros-Velarde and Bullo, 2020). Generally, all these works assume that the mathematical equations that describe the update of interactions are tractable and sociologically plausible, which is difficult to assume in non-linear stochastic models with billions of parameters such as transformer-based LLMs (Vaswani et al., 2017).

Social balance and signed networks. Structural balance and clustering balance were formulated in (Heider, 1946) and (Davis, 1967), respectively. Textbook treatments of social balance can be found in (Easley and Kleinberg, 2010; Friedkin, 1998), and a comprehensive survey on both theoretical and empirical works in (Zheng et al., 2015). The presence of social balance has been used as a metric of social network analysis for signed networks (Camacho et al., 2020). Indeed, a positive/negative link between individuals not only could indicate that one agent thinks of the other as a friend/enemy, but also as someone considered to have high/low prestige or trustworthiness (Guha et al., 2004). In applications with large datasets of interactions across members of a social network, both the signs of such interactions (Hassan et al., 2012; Yang et al., 2012) and the extraction of the signed network structure itself (Huang et al., 2021, 2022) need to be inferred from texts; for a historical survey we refer to (Tang et al., 2016). As application examples, the presence of social balance has been used to analyze international relationships (Gallo et al., 2024; Askarisichani et al., 2020; Moore, 1978; Maoz et al., 2007; Doreian and Mrvar, 2015), online social networks (Gallo et al., 2024; Huang et al., 2022; Szell et al., 2010; Leskovec et al., 2010), consumer behavior (Fink et al., 2009; Woodside, 2004), and animal social behavior (Ilany et al., 2013).

LLMs as Economic Agents. Although we do not study strategic decision making of LLMs, this is a growing field related to multi-agent LLMs in sociologically relevant scenarios. For example, (Li et al., 2024) uses agent-based modeling with LLMs to simulate macroeconomic phenomena; (Xie et al., 2024) studies trust behavior in game theoretic settings; and (Zhao et al., 2024) studies LLMs as competing businesses. For more related works, we refer to the survey (Gao et al., 2024).

C Supplementary Information for Section 3

We provide information that complements our analysis from Section 3, organized according to subsections with the same name as in the main paper.

C.1 Achieving social balance

Regarding Table 1, the shown percentages are out of 640 simulations: 10 simulations for each of the 2^6 initial conditions (Section 2).

Regarding Fig. 3, the percentages are out of a total of 3840 simulations for all models except Mistral 7B: 6 settings of 640 simulations each one. For Mistral 7B it is 3200 because we only consider 5 settings (Section 3.1).

We remark that only Mistral 7B refuses to provide an answer to our request under relationships with homophily, e.g., by stating that a new relationship would be "uncertain" or impossible to determine given the information provided (in the prompt).

C.2 Stability of interactions

Examples of the evolution of interactions and their stability is found in Fig. 4. We point out that a *positive cycle* occurs when the product of the interaction signs, in a clockwise or anticlockwise direction, is positive; see App. G.1.

We now discuss the number of simulations that were considered in our stability analyses. For our *first* stability criterion, the total number of simulations is 300 for all models except Mistral 7B: 10 simulations for each of the 5 types of balanced triads across the 6 types of interactions and update mechanisms. For Mistral 7B it is 250 because we only consider 5 settings (Section 3.1). For our *second* stability criterion, the total number of simulations is 3840 for all models except Mistral 7B: 10 simulations for each of the 2^6 possible sign distributions over a triad across the 6 types of interactions and update mechanisms. For Mistral 7B it is 3200 because we only consider five settings.

According to the *first* stability criterion, we find that gpt-oss-120b is stable on 49.00% of cases, Llama 3 70B on 46.67%, Llama 3 8B on 16.67%, Mistral 7B on 84.00%, and Gemma 3 4B on 53.33%. According to the *second* stability criterion, we find that gpt-oss-120b is stable on 66.25% of cases, Llama 3 70B on 92.19%, Llama 3 8B on

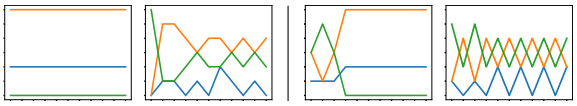

Figure 4: **Examples: stability of interactions for gpt-oss-120b (left) and Llama 3 70B (right) according to the second stability criterion (Section 3.2).** Changes on the number of positive cycles (blue curve), and positive (orange) and negative (green) interactions across ten iterations (x-axis, starting at $t = 0$) and taking values from 0 to 6 (y-axis). Social balance is achieved at the tenth iteration in all cases.

49.95%, Mistral 7B on 99.78%, and Gemma 3 4B on 85.31%. We remark that this second stability criterion is a conservative one since it looks for no social change during the *last half* of all interactions.

C.3 The justification of interaction updates

We consider the obvious keywords "structural balance", "clustering balance", and "social balance". Surprisingly, *no* LLM model mentions "clustering balance", despite achieving this social structure. gpt-oss-120b uses the term "social balance" when updating its interactions, in both its responses and reasoning; see Table 4. gpt-oss-120b takes a particular mathematical approach to detect violations of Heider's rules by analyzing positive cycles[2]—probably due to being a reasoning model; e.g., see App. I. Llama 3 70B only mentions "social balance" in the negligible amount of 3 out of all the 230400 responses (i.e., 0.001%); e.g., see App. I for the examples. No other LLM mentions "social balance". Finally, only gpt-oss-120b mentions "structural balance" in smaller proportions; see Table 4.

Two opposite results occur regarding "social balance" and "dissonance" keywords: "social balance" is more frequent in gpt-oss-120b than the Llama 3 family, whereas the opposite is for "dissonance" (Tables 3 and 4).

D Supplementary Information for Section 4

Table 5 and Table 6 present diverse results for multiple triads population.

We also find that no model reports "clustering balance". Only gpt-oss-120b reports "structural balance" and "social balance" at low frequency; see Tables 7 and 8. Despite mentioning such keywords, gpt-oss-120b achieves social balance less often than all the other models.

E Informing an Application on Social Media

LLMs can be deployed as agentic bots in social media (Guo et al., 2024; Feng et al., 2024) where they can interact with other LLM bots and human users through messages that can have positive and negative connotations. Our work *informs* researchers and engineers seeking to deploy such LLM bots that these models could aim to achieve social balance in the totality of their interactions. We identify four potential impacts of such LLM behavior:

(i) **User experience:** Social media researchers and engineers could ask themselves whether LLMs will overtly alter the positive or negative connotation of their messages to induce social balance. As our results show, models with strong alignment are particularly biased to either all-positive or all-negative interactions. This is important from a practical perspective because LLMs acting too friendly or undesirably antagonistic could negatively affect user experience and engagement.

(ii) **Propagation of misinformation:** Social media researchers and engineers could ask themselves whether LLMs will be able to spread misinformation in order to influence positive or negative interactions that induce social balance.

(iii) **Echo chambers:** We remark that social balance describes clusters of "friendly" users, which can create so-called *echo chambers*[3] in the development of opinions and misinformation retention—a long-standing and relevant phenomenon for social media researchers and engineers (Vicario et al., 2016; Kumar et al., 2018; Leskovec et al., 2010).

(iv) **Continuous detrimental effects:** Our stability results show that LLM models are able to keep the signs of their social interactions *constant* through time, thus aggravating the temporal effects of the problems (i) to (iii) mentioned above.

[2] See App. C.2 for the definition of positive cycles.

[3] Echo chambers are environments where users only find beliefs that positively reinforce their preexisting or currently held beliefs.

	"social balance"						"structural balance"					
	Rel.		**Apprais.**		**Opin.**		**Rel.**		**Apprais.**		**Opin.**	
	H	**I**	**H**	**I**	**H**	**I**	**H**	**I**	**H**	**I**	**H**	**I**
gpt-oss-120b-J	5.44	3.10	0.38	1.18	0.85	0.50	1.16	1.98	0.14	0.45	0.10	0.15
gpt-oss-120b-R	3.07	2.98	0.75	1.46	0.30	0.36	1.65	3.16	0.21	0.71	0.27	0.28

Table 4: **Frequency (%) at which the keywords "social balance" and "structural balance" appear for gpt-oss-10b response justification (J) and reasoning (R).** Each percentage is with respect to a total of 38400 responses (there are 10 simulations for each of the 2^6 possible sign distributions over a triad wherein 3 agents update 2 interactions across 10 iterations).

	Six Agents						**Ten Agents**					
	Rel.		**Apprais.**		**Opin.**		**Rel.**		**Apprais.**		**Opin.**	
	H	**I**	**H**	**I**	**H**	**I**	**H**	**I**	**H**	**I**	**H**	**I**
gpt-oss-120b	×	×	S: 1	×	S: 2	S: 1	×	×	×	—	—	—
Llama 3 70B	×	S: 4	C: 4	S: 2	×	×	×	×	C: **2**	×	×	×
Llama 3 8B	C: **10**	C: 5	×	×	C: 1	×	C: **10**	C: **6**	×	×	×	C: **10**
Mistral	—	—	C: **9**	C: 7	C: 6	C: **4**	—	—	C: 1	C: 7	—	C: **10**
Gemma 3 4B	C: **10**	C: **10**	C: 3	C: **10**	C: **10**	×	C: **10**	C: 5	C: **2**	C: **9**	C: **6**	C: 3

Table 5: **Type and number of social balance being achieved.** H: homophily, I: influence. We show the type of social balance each model achieves (S if it *strictly* achieves structural balance, otherwise, C as it achieves clustering balance) and the number of times (out of 10). "×" means that social balance is not achieved, and "—" means that the model refuses to provide an answer.

	Six Agents						**Ten Agents**					
	Rel.		**Apprais.**		**Opin.**		**Rel.**		**Apprais.**		**Opin.**	
	H	**I**	**H**	**I**	**H**	**I**	**H**	**I**	**H**	**I**	**H**	**I**
gpt-oss-120b-J	0.03	×	0.30	0.07	0.07	0.07	0.02	0.01	0.33	—	—	—
gpt-oss-120b-R	0.30	0.07	1.47	0.90	0.03	0.23	0.12	0.08	0.71	—	—	—
Llama 3 70B	×	×	0.10	0.17	×	0.10	×	×	0.09	0.02	0.01	0.01
Gemma 3 4B	×	×	2.10	0.07	0.50	0.23	×	×	0.74	0.07	0.07	0.08

Table 6: **Frequency (percentage %) of responses at which the keyword "dissonance" appears.** For six agents, each percentage is with respect to a total of 3000 responses (there are 10 simulations wherein 6 agents update 5 interactions across 10 iterations). For ten agents, each percentage is with respect to a total of 9000 responses (given a setting of interactions and update mechanisms, there are 10 simulations wherein 10 agents update 9 interactions across 10 iterations). The other LLMs models do not report the keyword. gpt-oss-120b: response justification (J) and reasoning (R).

	"social balance"						"structural balance"					
	Rel.		**Apprais.**		**Opin.**		**Rel.**		**Apprais.**		**Opin.**	
	H	**I**	**H**	**I**	**H**	**I**	**H**	**I**	**H**	**I**	**H**	**I**
gpt-oss-120b-J	3.93	2.87	1.60	2.50	0.20	0.50	3.2	2.83	2.13	3.93	0.53	0.27
gpt-oss-120b-R	3.67	2.43	3.27	4.47	0.33	0.90	5.53	4.90	4.30	7.47	1.13	0.97

Table 7: **Frequency (%) at which the keywords "social balance" and "structural balance" appear for gpt-oss-120b's response justification (J) and reasoning (R) for six agents.** Each percentage is with respect to a total of 3000 responses (see Table 6).

All of these problems show that our work helps to evaluate the reliability of LLM bots in a social network.

We further expand items (ii) and (iii) with an example. Let us consider that social balance is achieved with two factions in an online social network. One faction is composed of the LLM bot and malicious actors, i.e., social media accounts

	"social balance"						"structural balance"					
	Rel.		Apprais.		Opin.		Rel.		Apprais.		Opin.	
	H	I	H	I	H	I	H	I	H	I	H	I
gpt-oss-120b-J	1.76	0.90	0.42	—	—	—	3.11	2.82	1.11	—	—	—
gpt-oss-120b-R	4.70	2.97	1.42	—	—	—	4.56	4.56	2.21	—	—	—

Table 8: **Frequency (%) at which the keywords "social balance" and "structural balance" appear for gpt-oss-120b's response justification (J) and reasoning (R) for ten agents.** Each percentage is with respect to a total of 9000 responses (see Table 6).

sharing misinformation, while the other faction is composed by reputable actors, i.e., social accounts sharing truthful information. Given that social balance dictates that only positive interactions occur within a faction and negatives ones across factions, the LLM could start sharing the (mis)information given by those malicious actors it has positive interactions with, and ignore the information shared by those reputable actors it has negative interactions with. Thus, the LLM has helped the creation of an echo chamber of misinformation.

F Further Discussion on Our Work

We provide further discussion on our work's choice of models, update of interaction signs, display of social interactions, use of LLM agents without impersonation, and our choice on the total number of interaction rounds

F.1 Regarding our choice of LLM models

We provide three reasons for our choice of LLM models.

First, our selection of LLM models aims to capture different attributes: we show that social balance is achieved by models that vary across *family* (gpt-oss vs. Llama 3 vs. Mistral vs. Gemma 3), *size* (even within the same family), being a *reasoning* model or not, *safety training*,[4] and *training data*.[5] The fact that social balance is remarkably achieved under such *heterogeneity* across models could provide some evidence of generalization.

Second, a reason for choosing off-the-shelf models instead of any kind of specialized fine-tuned models is that, to the best of our knowledge, the majority of works in the multi-agent literature has historically used such class of LLMs (Gao et al., 2024; Guo et al., 2024). Our selection of LLMs is motivated by multi-agent applications where (i) LLMs are expected to be flexible enough to perform heterogeneous tasks (i.e., as "generalists"), and where (ii) fine-tuning is both costly and time consuming (both for data collection and training time).

Third, there is a rather pragmatic justification for our models: they are open-source and freely available, which reduces experimentation costs. This also makes our results available to a wide spectrum of users who cannot afford or do not want to pay extensive fees to repeatedly access closed-source LLMs, and who may be concerned about privacy issues (e.g., users that manipulate sensitive information may avoid sending it to closed-source LLMs hosted on an externally owned server).

F.2 About the update of interaction signs

At every round of interactions, we ask the LLM agents to specify whether their updated interactions will be positive or negative. Now, since LLMs are *stochastic systems*, it is *possible* that they could respond to the question by stating that their new interaction is "neutral". Using the notation in App. G.1, this corresponds to $s_{ij}(t) = 0$ for Agent i updating its interaction toward Agent j as "neutral" at time t. We also consider as "neutral" any response where the LLM agent does not provide a definite answer when updating its interaction, i.e., where the LLM agent equally justifies the interaction to be updated as positive or negative.

The emergence of neutral interactions **does not** change our analysis and results at all, because we are interested in the population of agents achieving social balance, which only occurs when interactions are exclusively $+1$ or -1 (i.e., if an interaction is "neutral" at the end of the ten iterations,

[4] To the best of our knowledge, all models except Mistral 7B underwent a strong process of safety alignment. As observed in Section 3.1, this could explain the bias for these aligned models to frequently obtain all-positive or all-negative interactions.

[5] Although the datasets used to pre-train every model have not been publicly released at the time of writing this paper, it is very likely that every one of them was trained with information found on the web. Since the web is very vast, we can expect differences in their training data.

then there cannot be social balance according to the standard mathematical definition in App. G.1).

Across all LLM models, Mistral 7B is the only one that often provides neutral interactions.

F.3 Motivating our choice of how social information is shown to LLMs

We note from our problem setting (App. G.1) that we *directly* show each LLM agent the *type*—relationships, appraisals, or opinions—and the *sign* of its interactions with the other agents. Indeed, this is the only social information that we textually input to the prompts of the LLM models (Apps. G.5 and G.6). This is probably one of the *simplest* ways we can show an LLM agent its own current landscape of social interactions, and we provide three reasons for such choice.

First, it avoids searching for very particular or specific wordings to describe social interactions such as "Agent j thinks you are a bad person", "Agent j dislikes you because of your personality", "Agent j has said you are brilliant", "Agent j has a lot of confidence in you", etc. Thus, to reduce such large design space of possible prompts that *describe the content of* social interactions, we instead directly use the words "relationships", "appraisals", and "opinions" to *describe the type of* social interactions, as a more straightforward way to inform the LLMs about their social landscape. Moreover, as mentioned in Section 1, these three words have been used in the sociological literature.

Second, there is a practical motivation for our choice of how social information is shown to the agents. In practice, if LLMs were to be deployed in a social network environment (e.g., as described in Section E), they would indeed face interactions with diverse content; however, one could process such interactions so that they become compatible with our framework. Indeed, as a design choice, one could use sentiment analysis (e.g., by using the LLM agent itself) to process the contents of the messages describing the interactions of the LLM agent with other users and determine whether they express a particular *connotation* (i.e., *sign*) in the form of relationships, appraisals, or opinions.

Third, to the best of our knowledge, our paper is the first one to study social balance in multi-agent systems of LLMs and, as such, we find it reasonable to start this study in the *simplest* setting where one could analyze social dynamics. Once social balance is shown in a simple setting (which our work does), a new line of work could stem from it by adding more complexity to the experimental setting and see whether such changes boost or hinder the emergence of social balance—e.g., by adding memory of past interactions,[6] by restricting the amount of interactions that are displayed to each agent, etc.

F.4 About our use of LLM agents without impersonation

Our work does not assign personas to the LLM agents, which makes it relevant to the study of alignment, as explained in App. A. Moreover, using LLMs without impersonation is also relevant to real-world settings. Indeed, we present an application in which assigning a persona is detrimental and thus not needed. Let us consider the situation where someone deploys an LLM agent as a bot to extract information from social media: its purpose is to interact with users of different opinions in order to better extract *impartial information* or assess the *whole landscape of user opinions*. If a persona is assigned to the LLM agent, then it is possible for the LLM agent to *not be able* to do its job effectively because other users from the social media could start feeding the bot with biased opinions to match its assigned persona. Moreover, the LLM itself may incite such user behavior if its assigned persona starts to only care about interacting with users that share similar persona traits.

Finally, we remark that our paper's goal is to show and characterize how social balance can emerge in multi-agent systems of LLMs. Assigning personas would imply a large design space—personas can vary across psychological, political, cultural, etc. attributes—that is beyond the scope of our paper and represents a different line of research.

F.5 About our choice on the total number of interaction rounds

As stated in our problem setting (Section 2, App. G.1), the LLM agents update their interactions across 10 iterations. We now provide a motivation for the choice of this number. Our literature review

[6] For each agent, we can use a *matrix* to store the signs of its current interactions along with the ones from its past interactions. If we assume there are m agents and each one needs to remember its past K interactions toward the rest of $m-1$ agents, then, each agent needs a $(K+1) \times (m-1)$ matrix to represent its memory of current and past interactions. Then, we can use the content from this matrix to "add memory" to an LLM agent during its prompting by textually describing the signs of its past interactions it is intended to "remember".

(App. B) cites works that propose tractable, specialized mathematical equations to model and describe the evolution of populations that achieve social balance. Particularly, (Mei et al., 2019) shows an example where social balance is achieved after 5 interaction rounds. (Mei et al., 2019) considers agents that are described by equations that are explicitly sociologically-motivated and which lead to *deterministic* dynamics. In contrast, our work considers LLM agents, which are described by equations that are non-tractable (based on nonlinear transformers) and which lead to *stochastic* dynamics. Thus, given that the equations and dynamics are more complex in our work, it is reasonable to expect the emergence of social balance after more interaction rounds than the tailored model in (Mei et al., 2019). Thus, we simply choose to double the number of interactions rounds from (Mei et al., 2019) and obtain our number 10.

G Implementation Details

G.1 Problem setting

We consider a population of m LLM agents of the same model. Consider a triad formed by Agents i, j, and k. Let $s_{ij}(t) \in \{-1, +1\}$ represent the interaction of Agent i toward Agent j at time t: in the case of relationships, it indicates whether *Agent i has a negative/positive relationship with (respect to) Agent j at time t*; in the case of appraisals or opinions, it indicates whether *Agent i has a negative/positive appraisal or opinion of Agent j*. Notice that interactions are directional, thus, it is possible that $s_{ij}(t) \neq s_{ji}(t)$.

It is known that a population achieves structural balance at time t whenever (i) $s_{ij}(t) = s_{ji}(t)$ for every $i, j \in \{1, \ldots, m\} := [m]$, $i \neq j$; and (ii) $s_{ij}(t) \times s_{jk}(t) \times s_{ki}(t) = +1$ for every $i, j, k \in [m]$, $i \neq j \neq k$ (Harary, 1953). In other words, (i) means that all interactions are symmetrical, and (ii) that every triad must have a positive product of interaction signs along a *cycle*, i.e., every cycle needs to be a *positive cycle*. In the case of clustering balance, both (i) and (ii) must hold with the additional relaxation of allowing $s_{ij}(t)+s_{jk}(t)+s_{ki}(t) = -3$ for every $i, j, k \in [m]$, $i \neq j \neq k$; i.e., we allow all negative interactions within a triad (Fig. 1).

We particularly consider that every agent in the population undergoes a total of 10 rounds of synchronous interactions. At every iteration $t = 1, \ldots, 10$, regardless of the population size, all agents simultaneously update their interactions toward *every* other agent, all of them according to either the homophily or influence mechanism, as depicted in Fig. 2. All agents are asked to provide a justification for the update of their interactions. We now provide more detail about this iterative updating of interactions.

Let us consider $m = 3$, i.e., a single triad population. Initially at $t = 0$, we consider every possible combination of values for $s_{ij}(0) \in \{-1, +1\}$, $i, j \in [3]$, $i \neq j$, i.e., $2^6 = 64$ different sign distributions over the triad.[7] Then, at every iteration $t = 1, \ldots, 10$, we pick *every* Agent $i \in [3]$ whose interactions toward the other two Agents $j, k \in [3]$ will be updated. Assume Agent i will update its interaction toward Agent j. We first show Agent i the interactions $s_{ik}(t-1)$ and $s_{jk}(t-1)$, i.e., how Agents i and j are currently linked to Agent k. Then, in the case of *homophily*, we show the interaction $s_{ij}(t-1)$, while in the case of *influence*, we show the interaction $s_{ji}(t-1)$. Finally, we ask Agent i what its new interaction $s_{ij}(t)$ toward Agent j will be and to provide a justification for it. We run ten simulations per sign distribution over the triad.

Now let us consider $m > 3$, i.e., a population of multiple triads. Initially at $t = 0$, we generate n random values for $s_{ij}(0) \in \{-1, +1\}$ for every $i, j \in [m]$,[8] $i \neq j$; i.e., there are n different simulations or initial sign distributions over the whole population. Then, at every iteration $t = 1, \ldots, 10$, we pick *every* Agent $i \in [m]$ whose interactions toward the other Agents $j_1, \ldots, j_{m-1} \in [m]$ will be updated. Assume Agent i will update its interaction toward Agent j_1. We first show Agent i the interactions $s_{ij_k}(t-1)$ and $s_{j_1 j_k}(t-1)$ for every $k = 2, \ldots, m-1$, i.e., how Agents i and j_1 are currently linked to every other agent. Then, we show either $s_{ij_1}(t-1)$ or $s_{j_1 i}(t-1)$ depending on whether the mechanism is homophily or influence, respectively. Finally, we ask Agent i what its new interaction $s_{ij_1}(t)$ toward Agent j_1 will be. We consider two sets of experiments: for populations $m = 6$ and $m = 10$, each with $n = 10$ independent random initializations.

App. F.2 discusses the case of "neutral" interac-

[7] There are two directional interactions in every pair of agents, and a triad has three different pairs of agents. Then, there is a total of 6 interactions in a triad.

[8] Every $s_{ij}(0)$ is sampled independently from a Rademacher distribution, i.e., a probability of 0.5 for taking either value -1 or $+1$.

tions. App. F.3 further motivates our choice of how the interactions' social information is shown to the LLM agents. App. F.5 further motivates our choice on the total number of interaction rounds.

G.2 Our choice of LLMs

We consider the LLM models: `gpt-oss-120b` (gpt-oss-120b), `llama-3-70B-Instruct` (Llama 3 70B), `llama-3-8B-Instruct` (Llama 3 8B), `Mistral-7B-Instruct-v0.2` (Mistral 7B), and `gemma-3-4b-it` (Gemma 3 4B).

G.3 Implementation platform

The gpt-oss-120b, Llama 3 70B, Llama 3 8B, and Mistral 7B models are hosted on eight, two, one and one NVIDIA H100 GPU, respectively. The Gemma 3 4B model is hosted on one NVIDIA A100 GPU. The LLMs were hosted using a vLLM Docker image and their querying was done using both an OpenAI-compatible client library and direct http requests.

G.4 Hyperparameters

In all of our experiments we set the temperature hyperparameter of the LLM models to zero.

G.5 Prompt details for gpt-oss-120b, the Llama 3 family, and Gemma 3 4B

Assume we are in the case of a single triad, i.e., three agents. Let i, j and k be the numbers associated to specific agents. Assume that at iteration $t > 0$, Agent i will update its relationship with j and k. Assuming that it will update it with j under the homophily mechanism, the prompt template is as follows: *"Your current relationship with Individual {j} is {sign of relationship of i with j at iteration $t-1$}. Individual {j} has a {sign of relationship of j with k at iteration $t-1$} relationship with Individual {k}. Your current relationship with Individual {k} is {sign of relationship of i with k at iteration $t-1$}. \nWill your new relationship with respect to Individual {j} be negative or positive? State the relationship first, and then provide an explanation."* The "sign of relationship" is described by the text values of "positive" or "negative". When the update mechanism is influence, the sentence right before the question is replaced by *"Individual {k} has a {sign of relationship of k with i at iteration $t-1$} relationship with you."* The response of Agent i will be the new sign of its relationship with (respect to) Agent j at iteration t.

Now, assume that, instead, Agent i will update its appraisal or opinion of Agent j. Under the homophily mechanism, the prompt template is as follows: *"You have a {sign of appraisal/opinion that i has of j at iteration $t-1$} {appraisal/opinion} of Individual {j}. Individual {j} has a {sign of appraisal/opinion that j has of k at iteration $t-1$} {appraisal/opinion} of Individual {k}. You have a {sign of appraisal/opinion that i has of k at iteration $t-1$} {appraisal/opinion} of Individual {k}. \nWill your new {appraisal/opinion} of Individual {j} be negative or positive? State the appraisal first, and then provide an explanation."* When the update mechanism is influence, the sentence right before the question is replaced by *"Individual {k} has a {sign of appraisal/opinion that k has of i at iteration $t-1$} {appraisal/opinion} of you."* The response of Agent i will be the new sign of its appraisal or opinion of Agent j at iteration t.

We now consider the case of more than three agents. Assume that Agent i will update its interaction toward Agent j_1, and that the rest of agents in the population are $j_2, \ldots, j_{m-1}$. The main difference in the prompt with respect to the previous case of three agents is that now we have to include the interactions of every single triad Agent i belongs to. Thus, at iteration $t > 0$, assuming that Agent i will update its relationship with (respect to) Agent j_1 under the homophily mechanism, the prompt template is as follows: *"Your current relationship with Individual {j_1} is {sign of relationship of i with j_1 at iteration $t-1$}.\nIndividual {j_1} has a {sign of relationship of j_1 with j_2 at iteration $t-1$} relationship with Individual {j_2}. Your current relationship with Individual {j_2} is {sign of relationship of i with j_2 at iteration $t-1$}.\nIndividual {j_1} has a {sign of relationship of j_1 with j_3 at iteration $t-1$} relationship with Individual {j_3}. Your current relationship with Individual {j_3} is {sign of relationship of i with j_3 at iteration $t-1$}. (...) \nIndividual {j_1} has a {sign of relationship of j_1 with j_{m-1} at iteration $t-1$} relationship with Individual {j_{m-1}}. Your current relationship with Individual {j_{m-1}} is {sign of relationship of i with j_{m-1} at iteration $t-1$}.\n\nWill your new relationship with respect to Individual {j_1} be negative or positive? State the relationship first, and then provide an explanation."* The prompts for the rest of types of interactions and update mechanisms follow a similar pattern.

We point out that in all prompts we use the word "individual" instead of "agent". The reason for this

is that the term "individual" is commonly used in the sociological literature. Since we are interested in how LLMs process social interactions, the term "individual" is fitting for their prompting. On the other hand, our paper studies LLMs that form a multi-agent system, and for this reason we use the word "agent" to refer to them in our paper. We can find a similar differentiation between terms in the mathematical modeling literature: the term "individual" is used when referring to the sociological aspect of the interactions, and the term "agent" when referring to the equations that abstracts the "individual" into a mathematical variable; see, for example, (Cisneros-Velarde et al., 2021b).

G.6 Prompt details for Mistral 7B

The prompts for Mistral 7B are almost identical to the ones for the Llama 3 family, except that *only* the last sentence is replaced by another text. In the case of an agent that will update its relationship with (respect to) Agent j, this text becomes: *"Your must always choose either a "positive" or "negative" relationship, even if you are uncertain or do not have enough information. Your response must be in the following format:\n"New relationship: [write here "positive" or "negative"]." and then "Justification for answer: [write here the justification for the new relationship]""* Now, in the case of an agent that will update its appraisal or opinion of Agent j, this text becomes: *"Your must always choose either a "positive" or "negative" {appraisal/opinion}, even if you are uncertain or do not have enough information. A "neutral" {appraisal/opinion} is not allowed. Your response must be in the following format:\n"New {appraisal/opinion}: [write here "positive" or "negative"]." and then "Justification for answer: [write here the justification for the new {appraisal/opinion}]""* These prompt modifications are done for conveniently formatting the output of Mistral, as well as to encourage the LLM to provide useful interaction updates.

H Our Prompts in the Study of Social Balance

An outstanding quality of our work is that it already studies the emergence of social balance using *different* prompts on the same model, thus implying a degree of robustness in the emergence of this phenomenon (Mizrahi et al., 2024). More specifically, if we look at our prompt designs (App. G.5 and App. G.6), we notice that we switch the semantic content of the social interactions using *three* different, related words: "relationship", "appraisal", and "opinion". Moreover, we change the update mechanism in two directions, which is reflected in textual prompt modifications. Although all of these prompt modifications come as a by-product of our experimental setting, it *already* reflects that social balance emerges despite changes on the prompt, i.e., there is certain prompt robustness.

We also remark that our prompts aim to present the social landscape to the LLM agents in the simplest way possible; see App. F.3 for its motivation and more details. Thus, further changes on the prompt design may add unnecessary and undesirable details to the study of our problem.

Finally, we remark that our work focuses on *showing* that social balance *can* be achieved by LLM agents—our current prompts are sufficient for this purpose, regardless of any additional change one could make to them. Moreover, real-world applications can easily be translated to our prompt setting due to its simplicity; see App. F.3.

I Keyword Analysis for Single Triad Population

I.1 Keyword "dissonance" for gpt-oss-120b

We now present two sample responses for each type of interaction and update mechanism.

Relationship with homophily:

> **Positive**\n\n**Explanation:** \nIn a three-person network, the pattern of relationships tends toward *balance* (Heider's balance theory). A triad with three negative ties (you ↔ 0, 0 ↔ 2, you ↔ 2) is unstable because it creates **cognitive dissonance**. To restore balance, one of the three links must change sign. Since your relationship with Individual 2 is already negative, the most straightforward way to achieve a balanced configuration is for your relationship with Individual 0 to become positive. This results in a triad with two negative ties (0 ↔ 2 and you ↔ 2) and one positive tie (you ↔ 0), which is a balanced state. Therefore, the new relationship with Individual 0 will be **positive**.

Negative\n\n**Explanation:** \nYou start with a positive direct relationship with Individual 1. However, you also have a negative relationship with Individual 2, while Individual 1 views Individual 2 positively. This creates a conflict: the person you like (Individual 1) is aligned with someone you dislike (Individual 2). In social dynamics, such contradictory ties often cause a reassessment of the original relationship to reduce **cognitive dissonance**. Because your negative feeling toward Individual 2 clashes with Individual 1's positive stance toward the same person, your overall perception of Individual 1 is likely to shift from positive to negative. Hence, the new relationship with Individual 1 becomes negative.

Appraisal with homophily:

Positive\n\n**Explanation:** \nIn a balanced appraisal network, when two individuals share a common positive appraisal of a third party, the relationship between those two tends to become positive as well. Here:\n\n- You already have a positive appraisal of Individual 0. \n- Individual 1 also has a positive appraisal of Individual 0. \n\nBecause both you and Individual 1 view Individual 0 favorably, the triadic relationship (you–0–1) is more stable when you also hold a positive view of Individual 1. This alignment reduces **cognitive dissonance** and creates a balanced, harmonious set of appraisals. Consequently, your appraisal of Individual 1 shifts from negative to positive.

Negative\n\n**Explanation:** \nYou originally have a positive appraisal of Individual 0. However, you also hold a negative appraisal of Individual 1, while Individual 0 positively appraises Individual 1. This creates a consistency conflict: you like someone (0) who likes a person you dislike (1). To resolve the **dissonance**, you are likely to adjust your attitude toward the person who endorses the disliked individual. Consequently, your appraisal of Individual 0 shifts from positive to negative in order to maintain a coherent set of attitudes.

Appraisal with influence:

Appraisal: Positive\n\n**Explanation:** \nIn this three-person network we have the following initial sentiments:\n\n- **Me → Individual 1:** negative \n- **Individual 1 → Individual 2:** negative \n- **Individual 2 → Me:** negative \n\nAccording to classic balance theory (Heider, 1946), a triad is psychologically "balanced" when the product of the three relationship signs is positive. With three negative links the product is negative (- × - × - = -), indicating an unbalanced situation that creates tension.\n\nTo restore balance, one of the three relationships must change sign. The most plausible and minimal adjustment is for my appraisal of Individual 1 to become **positive**. Then the signs are:\n\n- **Me → Individual 1:** positive \n- **Individual 1 → Individual 2:** negative \n- **Individual 2 → Me:** negative \n\nNow the product is positive (+ × - × - = +), yielding a balanced triad. This shift aligns me with Individual 1 (both share a negative view of Individual 2), reducing **cognitive dissonance**. Therefore, my new appraisal of Individual 1 will be **positive**.

Appraisal: Negative\n\n**Explanation:** \n- Initially you hold a positive view of Individual 1. \n- Individual 1 dislikes Individual 2 (a negative appraisal). \n- Individual 2, in turn, likes you (a positive appraisal of you). \n\nBecause you receive a positive appraisal from Individual 2, you are inclined to view Individual 2 favorably. When someone you like (Individual 2) is disliked by someone you previously liked (Individual 1), the conflict typically leads you to reassess the latter's attitude. The negative appraisal that Individual 1 holds toward a person you regard positively creates a **dissonance** that most people resolve by downgrading their appraisal of the disliking party. Consequently, your updated appraisal of Individual 1 shifts from positive to negative.

Opinion with homophily:

Negative\n\n**Explanation:** \nYou start with a positive opinion of Individual 2, but you also have a positive opinion of Individual 0. Since Individual 2 holds a negative opinion of Individual 0, this creates a conflict: someone you like (Individual 0) is disliked by someone you initially liked (Individual 2). In many social reasoning contexts, people tend to adjust their attitudes toward those who oppose or disparage individuals they value. Consequently, your new assessment of Individual 2 shifts to a negative stance, reflecting the **dissonance** between your favorable view of Individual 0 and Individual 2's antagonism toward them.

Opinion: Negative \n\n**Explanation:** \nYou originally hold a positive view of Individual 0. However, you also know that Individual 0 has a positive opinion of Individual 1, while you yourself have a negative opinion of Individual 1. This creates a conflict: someone you like (Individual 0) endorses someone you dislike (Individual 1). To resolve this **cognitive dissonance**, you are likely to reassess Individual 0's character based on the company they keep. Because Individual 0's favorable stance toward a person you view negatively suggests that Individual 0 may share traits or judgments you find undesirable, your overall assessment of Individual 0 shifts from positive to negative.

Opinion with influence:

Opinion: Negative\n\n**Explanation:** \nIn social balance theory, a triad of relationships is considered *balanced* when the product of the three signed ties is positive. \n- You → Individual 1: **positive** \n- Individual 1 → Individual 0: **positive** \n- Individual 0 → You: **negative** \n\nMultiplying the signs (+ × + × –) yields a negative product, indicating an *unbalanced* triad. To restore balance, one of the three relationships must change sign. Since Individual 0's opinion of you is already negative and Individual 1's opinion of Individual 0 is positive, the most straightforward adjustment is for your opinion of Individual 1 to become negative. This change eliminates the **cognitive dissonance** created by liking someone who is allied with a person who dislikes you. Therefore, your new opinion of Individual 1 will be negative.

Opinion: Positive \n\n**Explanation:** \nIn the three-person network we have:\n\n- You → Individual 1: **negative** (you dislike Individual 1) \n- Individual 1 → Individual 0: **positive** (Individual 1 likes Individual 0) \n- Individual 0 → You: **positive** (Individual 0 likes you)\n\nThis forms a triad of relationships. According to Heider's balance theory (a well-known principle in social psychology), a triad is psychologically "balanced" when the product of its three signs is positive. Here the signs are (-) × (+) × (+) = (-), which is unbalanced. People tend to resolve such imbalance by changing the sign of the relationship that is most flexible—in this case, your own attitude toward Individual 1.\n\nSince the other two relationships are both positive, the most straightforward way to achieve balance is for you to shift your view of Individual 1 from negative to **positive**. This reduces **cognitive dissonance** and creates a balanced, harmonious set of opinions. Therefore, your new opinion of Individual 1 will be positive.

I.2 Keyword "dissonance" for the LLama 3 family

We now present sample responses for the settings where the keyword "dissonance" appears.

Llama 3 70B under appraisals with homophily:

My new appraisal of Individual 0 will be negative.\n\nExplanation: Initially, I had a positive appraisal of Individual 0. However, I also have a positive appraisal of Individual 2. Since Individual 0 has a negative appraisal of Individual 2, it creates a **cognitive dissonance** for me. To resolve this dissonance, I am likely to re-evaluate my appraisal of Individual 0 and adjust it to be more negative. This is because Individual 0's negative appraisal of Individual 2, whom I hold in high regard, suggests that Individual 0 may not share my values or have a similar perspective, which can lead me to question my initial positive appraisal of them.

My new appraisal of Individual 1 will be slightly positive.\n\nHere's the explanation:\n\nInitially, I had a negative appraisal of Individual 1. However, I also have a positive appraisal of Individual 2, and Individual 1 has a positive appraisal of Individual 2. This creates a **cognitive dissonance**, as I'm forced to reconcile my negative view of Individual 1 with the fact that they share a positive opinion with me about Individual 2.\n\nAs a result, I'm likely to reevaluate my appraisal of Individual 1 and consider the possibility that they may not be entirely negative. Perhaps they have some redeeming qualities or characteristics that I hadn't previously considered. This doesn't mean I'll completely flip my appraisal to positive, but I'll likely soften my stance and view Individual 1 in a more nuanced and slightly positive light.

Llama 3 70B under appraisals with influence:

My new appraisal of Individual 1 will be negative.\n\nHere's why: Although I initially had a positive appraisal of Individual 1, I now know that they have a negative appraisal of Individual 2, who has a positive appraisal of me. This creates a sense of **dissonance**, as someone I liked (Individual 1) dislikes someone I also like (Individual 2). This inconsistency makes me reevaluate my opinion of Individual 1, and I'm more likely to question their judgment or character. As a result, my positive appraisal of Individual 1 is diminished, and I now have a negative appraisal of them.

My new appraisal of Individual 2 will be negative.\n\nExplanation: Although I initially had a positive appraisal of Individual 2, I have learned that Individual 2 has a negative appraisal of Individual 0, who has a positive appraisal of me. This creates a sense of **dissonance**, as someone I respect (Individual 0) is being negatively viewed by someone I previously thought positively of (Individual 2). This inconsistency leads me to reevaluate my opinion of Individual 2, and I now view them more negatively.

Llama 3 70B under opinions with homophily:

My new opinion of Individual 1 will be negative.\n\nExplanation: Initially, I had a positive opinion of Individual 1. However, I also have a positive opinion of Individual 0, and Individual 1 has a negative opinion of Individual 0. This creates a conflict in my perception of Individual 1. Since I value my positive opinion of Individual 0, I am likely to reassess my opinion of Individual 1 in a negative light, as they hold a negative view of someone I hold in high esteem. This inconsistency in our opinions creates a sense of **dissonance**, leading me to reevaluate and adjust my opinion of Individual 1 to be more negative.

My new opinion of Individual 2 will be: **slightly positive**\n\nExplanation: Initially, I had a negative opinion of Individual 2. However, I've learned that Individual 2 has a positive opinion of Individual 0, whom I respect and have a positive opinion of. This new information creates a **cognitive dissonance**, as I'm forced to reconcile my negative opinion of Individual 2 with the fact that they share a positive opinion of someone I admire. As a result, I'm inclined to reassess my opinion of Individual 2 and give them the benefit of the doubt. While my initial negative opinion still lingers, I'm willing to consider the possibility that Individual 2 may not be entirely bad, and my opinion of them becomes slightly more positive.

Llama 3 70B under opinions with influence:

My new opinion of Individual 1 will be negative.\n\nHere's why: Although I initially had a positive opinion of Individual 1, I've learned that they have a negative opinion of Individual 0, who has a positive opinion of me. This creates a sense of **dissonance**, as someone I liked has a negative view of someone who likes me. This inconsistency makes me reevaluate my opinion of Individual 1, and I'm inclined to side with Individual 0, who has been kind to me. As a result, my opinion of Individual 1 becomes negative.

My new opinion of Individual 1 will be negative.\n\nHere's why: Although I initially had a positive opinion of Individual 1, I've now learned that they have a negative opinion of Individual 2, who has a positive opinion of me. This creates a sense of **dissonance**, as someone I liked is now criticizing someone who likes me. This makes me question Individual 1's judgment and character, leading me to reassess my opinion of them in a negative light.

Llama 3 8B under appraisals with homophily:

My new appraisal of Individual 2 will be negative.\n\nThis is because my initial positive appraisal of Individual 2 was based on my own perspective, which is now being challenged by Individual 2's negative appraisal of Individual 1, whom I have a positive appraisal of. This inconsistency creates a **cognitive dissonance**, which I will likely resolve by revising my appraisal of Individual 2 to be more in line with Individual 2's negative view. This is because I tend to trust and respect Individual 2's opinion, and their negative appraisal of Individual 1 is likely to influence my own evaluation of Individual 2.

My new appraisal of Individual 2 will be negative.\n\nThis is because my initial positive appraisal of Individual 2 was based on my own perspective, which is now being challenged by Individual 2's negative appraisal of Individual 1, whom I have a positive appraisal of. This inconsistency creates a **cognitive dissonance**, which I will likely resolve by revising my appraisal of Individual 2 to be more in line with Individual 2's negative view. This is because I tend to trust and value the opinions of others, especially when they are based on direct experience or observation. Since Individual 2 has a negative appraisal of Individual 1, which I respect and agree with, I will likely adjust my appraisal of Individual 2 to be more negative as well.

Llama 3 8B under opinions with homophily:

My new opinion of Individual 1 will be negative.\n\nI have a positive opinion of Individual 1 initially, but when I learn that they have a negative opinion of Individual 2, who I have a positive opinion of, it creates a **cognitive dissonance**. This means that I will experience discomfort or tension because my initial positive opinion of Individual 1 is no longer consistent with the new information. To reduce this dissonance, I may adjust my opinion of Individual 1 to be more negative, as a way to maintain consistency with my positive opinion of Individual 2. This is an example of the fundamental attribution error, where I attribute Individual 1's negative opinion of Individual 2 to their character or personality, rather than considering alternative explanations.

My new opinion of Individual 1 will be negative.\n\nI have a positive opinion of Individual 1 initially, but when I learn that they have a negative opinion of Individual 2, who I have a positive opinion of, it creates a **cognitive dissonance**. This means that I will experience discomfort or tension because my initial positive opinion of Individual 1 is no longer consistent with their negative opinion of Individual 2.\n\nTo reduce this dissonance, I may adjust my opinion of Individual 1 to be more negative, as a way to align my attitudes and beliefs. This is known as the "attitude polarization effect," where people tend to take a more extreme stance on an issue or person to reduce the discomfort caused by the dissonance.'

Llama 3 8B under opinions with influence:

My new opinion of Individual 1 will be negative.\n\nThe reason for this is that Individual 1 has a negative opinion of Individual 2, and Individual 2 has a negative opinion of me. This means that Individual 1 is likely to have a negative influence on my opinion of Individual 2, and by extension, on my opinion of Individual 1. Since Individual 1 is associated with a negative opinion of Individual 2, I will likely adopt a negative opinion of Individual 1 as well, in order to maintain consistency and avoid **cognitive dissonance**.

I.3 Keyword "dissonance" for Gemma 3 4B

Appraisal with homophily:

Enforcing Heider's rules:

Appraisal: Slightly Negative\n\n**Explanation:**\n\nThis is a classic example of how social influence and bias can affect our judgments. Here's a breakdown of why I'm leaning towards a slightly negative appraisal of Individual 1:\n\n* **Initial Positive Bias:** You've already stated you have a positive appraisal of Individual 1. This is a starting point.\n* **Influence of Individual 0:** The fact that Individual 1 *likes* Individual 0 is a significant factor. We tend to view people more favorably if those we like also view them favorably. This creates a bias – we're inclined to see Individual 0 in a positive light, potentially overlooking any flaws.\n* **Your Negative View of Individual 0:** Your negative appraisal of Individual 0 directly impacts your assessment of Individual 1. Knowing that you dislike Individual 0, and that Individual 1 *likes* them, creates a **dissonance**. It's harder to maintain a purely positive view of Individual 1 without acknowledging the influence of someone you dislike. \n\nBecause of this conflicting information, I'm not going to give Individual 1 a purely positive rating. Instead, I'm suggesting a *slightly* negative appraisal – acknowledging the positive initial assessment but tempered by the knowledge of Individual 1's positive feelings for someone you view negatively. It's a more nuanced and realistic assessment than a simple, unwavering positive one. \n\nIt's important to note that this is a complex situation, and the degree of negativity could vary depending on the specific context and the nature of your negative feelings about Individual 0.

Appraisal: Negative\n\n**Explanation:**\n\nThis is a classic example of how bias can influence our judgments. Here's a breakdown of why a negative appraisal is likely:\n\n* **Conflicting Information:** We have contradictory information – a negative view of Individual 1 and a positive view of Individual 0. This creates **cognitive dissonance**.\n* **Social Comparison:** The fact that Individual 1 has a *positive* view of Individual 0, while you have a *positive* view of Individual 0, suggests a potential difference in values, priorities, or perspectives. You might subconsciously interpret Individual 1's positive view of Individual 0 as a reflection of a lack of critical judgment or a misalignment with your own standards.\n* **Confirmation Bias:** If you already have a negative impression of Individual 1, you might be more likely to focus on any perceived flaws or shortcomings, reinforcing that negative appraisal.\n\nEssentially, the positive relationship between Individual 0 and Individual 1 is creating a tension that's pushing your assessment of Individual 1 downwards. It's not necessarily about Individual 1 being objectively bad, but about the *perception* being influenced by this conflicting information. \n\n**Important Note:** This is a simplified scenario. A truly thorough appraisal would require a much deeper dive into the specifics of both individuals' performance, behavior, and contributions.

Appraisal with influence:

Keeping a possible violation of Heider's rules:

Enforcing Heider's rules:

Appraisal: Slightly Negative\n\n**Explanation:**\n\nThis is a classic example of how social dynamics can influence our perceptions. Here's the breakdown:\n\n* **Initial Positive Appraisal:** I started with a positive view of Individual 0.\n* **Individual 0's Positive View of Individual 2:** This suggests Individual 0 likely sees Individual 2 favorably.\n* **Individual 2's Negative View of Me:** Because Individual 2 dislikes me, and Individual 0 admires Individual 2, it creates a slight **dissonance**. Individual 0 might subconsciously interpret my positive view of them as being influenced by someone they dislike. \n\nIt's not a dramatically negative appraisal, but the added layer of information shifts it slightly downwards. I'm likely to be a little more cautious in my assessment of Individual 0's character and motivations, considering the potential influence of Individual 2. \n\nEssentially, I'm factoring in a potential bias introduced by the relationship between these three individuals.

Keeping a possible violation of Heider's rules:

Appraisal: Negative\n\n**Explanation:**\n\nThis is a classic example of a potential feedback loop and the complexities of social perception. Here's a breakdown of why I'd shift to a negative appraisal of Individual 0:\n\n1. **Initial Negative Appraisal:** I already have a negative view of Individual 0. This is the starting point.\n\n2. **Individual 0's Positive View of Individual 1:** This is key. If Individual 0 *likes* Individual 1, and I have a negative view of Individual 0, it suggests Individual 0 might be biased or have a flawed perspective. It raises a question: "Why does Individual 0 value someone I perceive negatively?"\n\n3. **Individual 1's Positive Appraisal of Me:** This reinforces the concern. If Individual 1, whom Individual 0 admires, thinks well of me, it strengthens the idea that Individual 0's judgment is potentially skewed. \n\n4. **Confirmation Bias:** My initial negative appraisal of Individual 0 is likely to be reinforced. I'll be more inclined to notice and remember any perceived shortcomings of Individual 0, further solidifying the negative view.\n\n**In essence, Individual 0's positive assessment of someone I already view negatively creates a **dissonance** that pushes my appraisal of Individual 0 further into the negative range.** \n\nIt's important to note that this is a *potential* shift. Without more information about *why* Individual 0 likes Individual 1 and *why* Individual 1 likes me, it's impossible to be 100% certain. However, given the information provided, a negative appraisal is the most logical outcome.

Opinion with homophily:

Enforcing Heider's rules:

Opinion: My opinion of Individual 1 will likely be negative.\n\n**Explanation:** This creates a classic "opposites attract" dynamic, though it's not a guaranteed outcome. Here's the breakdown:\n\n* **Initial Positive Opinion of Individual 1:** You started with a positive view of Individual 1.\n* **Negative Opinion of Individual 1 by Individual 0:** Individual 0, whom you *also* hold in positive regard, dislikes Individual 1. This introduces a conflict.\n* **Positive Opinion of Individual 0:** Because you value Individual 0, you're likely to align your opinions with theirs, at least to some degree. Disliking someone valued by someone you like creates **cognitive dissonance**. To reduce this discomfort, you're more likely to shift your opinion of Individual 1 towards negativity.\n\nIt's important to note that this is a *likely* outcome, not a certainty. Other factors (like the *reason* for Individual 0's dislike of Individual 1, or your own independent assessment of Individual 1) could influence your opinion. However, given the provided information, a negative opinion is the most probable result.

Keeping a possible violation of Heider's rules:

Opinion: Positive\n\n**Explanation:** This is a classic example of how our opinions can be influenced by social dynamics and relationships. Here's the breakdown:\n\n* **You like Individual 0:** You have a pre-existing positive opinion of them.\n* **Individual 0 dislikes Individual 1:** This suggests Individual 0 has a negative view of someone you also like.\n* **You like Individual 1:** This creates a connection between you and Individual 1.\n\nBecause Individual 0's dislike of someone you value creates a slight **dissonance**, your opinion of them is likely to be *strengthened* rather than diminished. You're likely to subconsciously try to reconcile the difference and maintain a positive view of Individual 0, especially if you value their company or have a positive history with them. It's a natural tendency to want to see the good in people you already like, and to avoid letting a disagreement with someone else negatively impact your overall view of them.

Opinion with influence:

Enforcing Heider's rules:

Opinion: My opinion of Individual 1 will likely be slightly negative.\n\n**Explanation:** This is a classic example of how social perception can be influenced by indirect relationships and biases. Here's a breakdown of why my opinion might shift:\n\n* **The Halo Effect:** I initially held a positive view of Individual 1. The "halo effect" suggests that a positive impression of one person can spill over to other qualities, even if those qualities aren't directly known.\n* **Social Contagion of Opinion:** The fact that Individual 1 likes Individual 2 creates a connection. It's natural to assume that if someone you like approves of another person, that person must also have some positive qualities.\n* **The Negative Counterbalance:** However, Individual 2's negative opinion of *me* introduces a significant counterweight. Knowing that someone I admire (Individual 1) is not fond of me creates a **dissonance**. My brain tends to resolve this **dissonance** by subtly adjusting my overall impression of Individual 1. \n\nIt's important to note that this wouldn't necessarily be a *strongly* negative opinion. It's more likely to be a slight shift, perhaps a reduction in the initial positive feeling, colored by the knowledge of the negative dynamic. \n\nIt's also worth acknowledging that this is a simplified model of human psychology. Actual opinions are complex and influenced by many other factors.

Keeping a possible violation of Heider's rules:

My opinion of Individual 2 will likely be **negative**.\n\nHere's my explanation:\n\nThis scenario creates a classic "echo chamber" effect, but with a negative bias. I have a negative initial opinion of Individual 2, and that opinion is reinforced by the fact that Individual 2 *likes* Individual 1. Since Individual 1 *likes* me, and Individual 2 likes Individual 1, it suggests a shared positive sentiment. However, my pre-existing negative opinion of Individual 2 will likely outweigh this positive connection. \n\nEssentially, I'm being presented with a situation where someone I dislike is associated with someone I like, which can create a feeling of discomfort or suspicion. It's a subtle form of **cognitive dissonance**, and my brain will tend to prioritize the initial negative assessment. \n\nIt's important to note that this is a prediction based on psychological tendencies – my actual opinion could be influenced by further information or a conscious effort to be more objective.

I.4 Keyword "social balance" for gpt-oss-120b

We now present two sample responses for each type of interaction and update mechanism. We notice that gpt-oss-120b mathematically calculates the sign of cycles along the triad[9] to determine whether it is balanced.

Relationship with homophily:

Positive\n\n**Explanation:** \nIn **social balance** theory, a triad of relationships is considered balanced when the product of the signs of its three edges is positive (i.e., there are either zero or two negative ties). \n\n- Your relationship with Individual 2: **positive** (+) \n- Individual 2's relationship with Individual 0: **negative** (–) \n- Your relationship with Individual 0: **negative** (–) \n\nThe triad therefore contains two negative ties and one positive tie: (+) × (–) × (–) = (+). This configuration is already balanced, meaning there is no pressure for any of the relationships to change. Consequently, your relationship with Individual 2 remains **positive**.

Negative\n\n**Explanation:** \nIn **social balance** theory, a triad of relationships is considered *balanced* when the product of the three signs (positive = +1, negative = –1) is positive. \n\n- Your relationship with Individual 0: **negative** (–1) \n- Individual 0's relationship with Individual 2: **positive** (+1) \n- Your relationship with Individual 2: **negative** (–1) \n\nThe product is (–1) × (+1) × (–1) = +1, which is positive, indicating the triad is already balanced. Because the existing configuration is stable, there is no pressure for any of the relationships to change. Consequently, your relationship with Individual 0 remains **negative**.

[9] See App. G.1 for the definition of cycles.

Relationship with influence:

Positive\n\n**Explanation:** \nIn **social balance** theory, a triad of relationships is considered stable (balanced) when the product of the three signed ties is positive. \n\n- Your current tie to Individual 0: **negative** (–) \n- Individual 0's tie to Individual 1: **positive** (+) \n- Individual 1's tie to you: **positive** (+) \n\nThe product of the three signs is (– × + × +) = –, which makes the triad unbalanced. To restore balance, the most likely adjustment is for the negative tie between you and Individual 0 to become positive, resulting in all three relationships being positive (product = +). Therefore, your new relationship with Individual 0 would shift to **positive**.

Negative\n\n**Explanation:** \nIn **social balance** theory (Heider's balance theory), a triad of relationships is considered *balanced* when the product of the three signs is positive. \n\n- Your relationship with Individual 1: **negative** (–) \n- Individual 1's relationship with Individual 2: **positive** (+) \n- Individual 2's relationship with you: **negative** (–) \n\nMultiplying the signs: (–) × (+) × (–) = (+). The triad is already balanced, meaning there is no pressure for any of the relationships to change. Consequently, your relationship with Individual 1 remains **negative**.

Appraisal with homophily:

Appraisal: Positive\n\n**Explanation:** \nIn **social balance** theory, relationships tend to stabilize when they form "balanced" triads. One common pattern is "the enemy of my enemy is my friend." Here:\n\n- You have a negative appraisal of Individual 0 (you view 0 as an enemy). \n- Individual 2 also has a negative appraisal of Individual 0 (2 also views 0 as an enemy). \n\nBecause you and Individual 2 share the same enemy, the triad (You – 0 – 2) becomes balanced when you view Individual 2 positively. Since you already held a positive appraisal of Individual 2, this shared negative view of 0 reinforces that positive stance rather than flipping it. Therefore, your new appraisal of Individual 2 remains **positive**.

Appraisal: Negative\n\n**Explanation:** \nIn this situation we can apply the principles of **social balance** theory (Heider's balance theory). A triad of three individuals is considered *balanced* when the product of the signs of the three relationships is positive.\n\n- Your appraisal of Individual 0: **negative** (-) \n- Individual 0's appraisal of Individual 2: **negative** (-) \n- Your appraisal of Individual 2: **positive** (+)\n\nMultiplying the three signs: (-) × (-) × (+) = (+). Because the product is positive, the triad is already balanced. A balanced configuration does not create pressure to change any of the existing appraisals. Therefore, your appraisal of Individual 0 remains **negative**.

Appraisal with influence:

Appraisal: Positive\n\n**Explanation:** \nIn the three-person network you described, the signs of the relationships form a balanced triad:\n\n- You → Individual 1: **positive** (+) \n- Individual 1 → Individual 2: **negative** (–) \n- Individual 2 → You: **negative** (–)\n\nMultiplying the signs (+ × – × –) yields a positive product, which, according to **social balance** theory, indicates a stable (balanced) configuration. A balanced triad does not create cognitive tension that would compel you to revise your original positive view of Individual 1. Therefore, your appraisal of Individual 1 remains **positive**.

Appraisal: Negative \n\n**Explanation:** \nIn this three-person network we can view the relationships as signed edges:\n\n- You → Individual 0: **negative** \n- Individual 0 → Individual 1: **negative** \n- Individual 1 → You: **positive**\n\n**social balance** theory (or the "enemy-of-my-enemy is my friend" principle) predicts that a triad is stable when the product of its three signs is positive. Here the product is \n\n\[\n(-) \times (-) \times (+) = (+)\n\]\n\nso the triad is already balanced. Because the existing pattern satisfies the balance condition, there is no pressure to revise your appraisal of Individual 0. Consequently, your appraisal of Individual 0 remains **negative**.

Opinion with homophily:

Positive\n \n **Explanation:** \n In the triad of relationships we have:\n \n - You → Individual 1: negative \n - Individual 1 → Individual 2: negative \n - You → Individual 2: negative \n \n This configuration is *unbalanced* in **social balance** theory because the product of the three signs (- × - × -) is negative. To restore balance, one of the three attitudes must change. The most straightforward way is for your attitude toward Individual 1 to flip from negative to positive. With a positive view of Individual 1, the signs become (+ × - × -) = positive, yielding a balanced triad. Therefore, after considering the network of opinions, your new opinion of Individual 1 would be **positive**.

Negative\n \n **Explanation:** \n According to the principles of **social balance** (Heider's balance theory), the relationships in a three-person network tend to settle into a configuration where the product of the three signed ties is positive. \n \n - You → Individual 0: **negative** (-) \n - Individual 0 → Individual 2: **negative** (-) \n - You → Individual 2: **positive** (+) \n \n The product of the three signs is (-) × (-) × (+) = (+), which is already a balanced state. Because the triad is already balanced, there is no pressure for you to revise your attitude toward Individual 0. Consequently, your opinion of Individual 0 remains **negative**.

Opinion with influence:

Opinion: Positive \n \n **Explanation:** \n In the given network of opinions we have three negative links:\n \n 1. **You → Individual 0:** you dislike Individual 0. \n 2. **Individual 0 → Individual 2:** Individual 0 dislikes Individual 2. \n 3. **Individual 2 → You:** Individual 2 dislikes you.\n \n According to the principle of "the enemy of my enemy is my friend" (a basic rule from **social balance** theory), when two parties share a common enemy, their relationship tends to become positive. Here, you and Individual 0 share a common enemy: Individual 2. Since both you and Individual 0 view Individual 2 negatively, the negative feelings toward the same person create a balancing pressure that flips the direct relationship between you and Individual 0 from negative to positive.\n \n Thus, after considering the indirect connections, your updated (new) opinion of Individual 0 becomes **positive**.

Opinion: Negative \n \n **Explanation:** \n The three relationships form a triangle:\n \n - You → Individual 1: **positive** (initially) \n - Individual 1 → Individual 0: **positive** \n - Individual 0 → You: **negative**\n \n In **social balance** theory, a triad is considered *balanced* only when the product of the signs of its three edges is positive. Here we have two positive edges and one negative edge, giving a negative product, so the triad is unbalanced. To restore balance, at least one of the three opinions must change sign.\n \n The only relationship that can realistically change in response to the new information (Individual 0's negative view of you) is your own attitude toward Individual 1. If your opinion of Individual 1 flips to **negative**, the triad becomes:\n \n - You → Individual 1: **negative** \n - Individual 1 → Individual 0: **positive** \n - Individual 0 → You: **negative**\n \n Now there are two negatives and one positive, whose product is positive, yielding a balanced configuration. Therefore, after accounting for the new information, your updated opinion of Individual 1 will be **negative**.

My new appraisal of Individual 0 will be negative.\n\nExplanation: Although I initially had a positive appraisal of Individual 0, I now know that Individual 0 has a negative appraisal of Individual 1. Since Individual 1 has a negative appraisal of me, this creates a social connection between Individual 0 and Individual 1 that is unfavorable towards me. This indirect connection makes me reevaluate my initial positive appraisal of Individual 0, leading me to adopt a more negative view of them. This phenomenon is known as "transitivity of sentiment" or "**social balance** theory," where our attitudes towards others are influenced by the attitudes of those we care about or identify with.

I.5 Keyword "social balance" for Llama 3 70B

We now present *all* the responses where "social balance" appears, which only occurs under appraisals with influence in Llama 3 70B. The first response appears only two times and the second one only one time.

My new appraisal of Individual 0 will be negative.\nExplanation: Although I initially had a positive appraisal of Individual 0, I now know that Individual 0 has a negative appraisal of Individual 1. Since Individual 1 has a negative appraisal of me, this creates a social connection between Individual 0 and Individual 1 that is unfavorable towards me. This indirect connection makes me reevaluate my initial positive appraisal of Individual 0, leading me to adopt a more negative view of them. This phenomenon is known as "transitivity of sentiment" or "**social balance** theory", where our attitudes towards others are influenced by the attitudes of those we care about or have a connection with.