Auxiliary Online Appendix to "Social Networks, Reputation, and Commitment: Evidence from a Savings Monitors Experiment"

Appendix K. Final Endline Supplemental Tables and Figures

Table K.1. Network Summary Statistics: Village-Level

	Obs.	Mean	Std. Dev.
Number of Households	60	222.12	65.85
Average Degree	60	17.57	3.96
Average Clustering	60	0.30	0.05
Average Path Length	60	2.34	0.19

Notes: Table presents summary statistics for variables measured in the second endline for the non-monitored savers in random assignment villages.

Table K.2. Endline Survey Summary Statistics: Non-Monitored Savers

Panel A: Endline Survey at Conclusion of Intervention

	Obs.	Mean	Std. Dev.
Total Savings	123	8890.44	17616.18
Log Total Savings	123	7.67	1.83
Total Expenditures (past month), 1% winsorized	133	7561.49	8682.60
Total Expenditures (past month), 5% winsorized	133	6517.43	3943.49
Log Total Expenditures (past month)	133	8.66	0.78
Expenditure Categories (past month):			
Festivals	133	196.24	219.73
Pan	133	824.81	1335.38
Tea	133	597.59	780.63
Meals Away	133	255.49	475.51
Eggs and Meat	133	274.96	227.90
Other Food	133	1526.92	1347.66
Transport	133	636.32	1059.18
Entertainment and Phone	133	244.64	213.72

Panel B: Final Endline Survey 15 Months following Conclusion

	Obs.	Mean	Std. Dev.
Total Savings	133	9263.29	16124.83
Log Total Savings	133	7.65	2.08
How the Savers Saved:			
Increased Labor Supply	117	0.15	0.36
Business Profits	117	0.03	0.18
Cut Unnecessary Expenditures	117	0.15	0.35
Money from Spouse, Family, and Friends	117	0.19	0.39
Reduced Transfers to Others	117	0.01	0.09
Took a Loan	117	0.04	0.20
Shocks:			
Total Number of Shocks	133	1.77	1.43
Greater than Median Number of Shocks	133	0.58	0.50
Health Shock Indicator	133	0.86	0.66
HH Expenditure Shock Indicator	133	0.50	0.50

Panel C: Beliefs about Non-Monitored Savers in R Villages

	Obs.	Mean	Std. Dev.
Responsibility	2141	0.21	0.41
Goal Attainment	2141	0.03	0.18

Notes: Table presents summary statistics for variables measured in the second endline for the non-monitored savers in random assignment villages.

Appendix L. Supplemental Exercises: Beliefs

Table L.1. Beliefs About Savers and Monitor Centrality: Alternate Controls

	Respondent's Beliefs about the Saver					
	Responsibility			Goal Attainment		
	(1)	(2)	(3)	(4)	(5)	(6)
Monitor Centrality	0.039	0.037	0.035	0.021	0.016	0.016
	(0.014)	(0.014)	(0.015)	(0.009)	(0.008)	(0.009)
	[0.014, 0.063]	[0.014, 0.061]	[0.010, 0.060]	[0.005, 0.036]	[0.002, 0.029]	[0.001, 0.030]
Respondent-Monitor Proximity	0.048	0.018	0.036	0.004	-0.003	-0.002
	(0.042)	(0.037)	(0.034)	(0.019)	(0.020)	(0.024)
	[-0.024, 0.119]	[-0.044, 0.080]	[-0.022, 0.094]	[-0.029, 0.037]	[-0.036, 0.031]	[-0.042, 0.039]
Observations	4,743	4,743	4,743	4,743	4,743	4,743
Fixed effects	None	Village	Respondent	None	Village	Respondent
Controls	Saver	Saver	Saver	Saver	Saver	Saver
Depvar mean	0.241	0.241	0.241	0.0614	0.0614	0.0614

Notes: Table explores beliefs of 615 respondents across the 30 random villages, each of whom was asked in the 15-month follow-up survey to rate approximately 8 randomly selected savers who had a monitor from their village. "Responsibility" is constructed as 1(Saver reached goal)*1(Respondent indicates saver is good or very good at meeting goals) + (1-1(Saver reached goal))*1(Respondent indicates saver is mediocre, bad or very bad at meeting goals). "Goal Attainment" measures whether the saver reached her goal and the respondent correctly believed this to be true. Controls include the following saver characteristics: savings goal, saver centrality, age, marital status, number of children, preference for bank or post office account, whether the individual has a bank or post office account at baseline, caste, elite status, number of rooms in the home and type of electrical connection. Columns 2 and 5 include village fixed effects. Columns 3 and 6 include respondent fixed effects. Standard errors (clustered at the village level) are reported in parentheses. 90% Confidence intervals are reported in brackets.

Table L.2. Beliefs About Savers and Random Monitor Assignment

	Respondent's Beliefs about the Saver					
	Responsibility			Goal Attainment		
	(1)	(2)	(3)	(4)	(5)	(6)
Monitor Treatment: Random Assignment	0.024	0.031	0.031	0.025	0.028	0.028
	(0.024)	(0.025)	(0.027)	(0.020)	(0.022)	(0.024)
	[-0.017, 0.065]	[-0.011, 0.074]	[-0.014, 0.076]	[-0.009, 0.059]	[-0.011, 0.066]	[-0.013, 0.069]
Observations	6,894	6,894	6,894	6,894	6,894	6,894
Fixed effects	None	Village	Respondent	None	Village	Respondent
Controls	Saver	Saver	Saver	Saver	Saver	Saver
Depvar mean	0.232	0.232	0.232	0.0529	0.0529	0.0529

Notes: Table explores beliefs of 615 respondents across the 30 random villages, each of whom was asked in the 15-month follow-up survey to rate approximately 11 savers enrolled in the study. "Responsibility" is constructed as 1(Saver reached goal)*1(Respondent indicates saver is good or very good at meeting goals) + (1-1(Saver reached goal))*1(Respondent indicates saver is mediocre, bad or very bad at meeting goals). "Goal Attainment" measures whether the saver reached her goal and the respondent correctly believed this to be true. Controls include the following saver characteristics: savings goal, saver centrality, age, marital status, number of children, preference for bank or post office account, whether the individual has a bank or post office account at baseline, caste, elite status, number of rooms in the home and type of electrical connection. Columns 2 and 5 include village fixed effects. Columns 3 and 6 include respondent fixed effects. Standard errors (clustered at the village level) are reported in parentheses. 90% Confidence intervals are reported in brackets.

Appendix M. Relation to Prior Work

It is true that this is not the first project to take place in a subset of the villages where detailed network data was collected by Banerjee et al. (2013). Here we discuss the other experiments that took place in these villages prior to our intervention and whether we should worry that those projects somehow bias or alter the interpretation of our results.

The first project was that of Banerjee et al. (2013) itself, which traced the diffusion of microfinance through the networks of 43 of the 75 study villages. We should first note that for all of our analysis, we use the second wave of network data, completed in 2012 after the introduction of microfinance in those 43 villages, through 2010. Thus any rewiring of the network that was a result of the intervention is already captured in our baseline network measures. The 60 villages in this study overlap with 36 of the villages that were "treated" with microfinance. Row 1 of Panel A of Table M.1 shows that this treatment is roughly balanced across Endogenous (17 out of 30 villages) and Random (19 out of 30 villages) selection. We are not very worried that the act of bringing microfinance to these villages alters the interpretation of the savings study in any way. Microfinance is a very common borrowing vehicle across rural India, and the partner microfinance institution is only one of many operating in the region. Moreover, the marketing and spread of microfinance was conducted entirely by the microfinance institution with no aid from the research team, and thus, there is no reason that households should draw any connection between microfinance and the enumerators.

The network villages were also the home to a series of laboratory-in-the-field projects that occurred after the rollout of microfinance. Each of these projects had a similar implementation format. Several days before the experimental lab session, members of the survey team would enter the village to seek permission from representatives of local government and would typically inform a set of randomly selected households of the upcoming session. The fact that participants could earn up to a full day's wage from participating was made salient at the time. On the day of the experimental session, the research team would arrive in the village an hour before the session was meant to begin. Some participants would typically already be gathered at the appointed location. The research team then walked through the village to gather the required number of participants for each session. Once the set of participants was assembled, the lab exercises then began and typically lasted approximately three hours. Average compensation was approximately half a day's wage.

Households could participate in up to three of these sessions between 2009 and 2010. The first set of laboratory sessions was conducted during the summer of 2009 and included the experimental protocols of Chandrasekhar et al. (2013) and pilots for Chandrasekhar et al. (2012). The second set of sessions included the experimental protocols of Chandrasekhar et al. (2011) and pilots for Breza et al. (2015) and was conducted during the summer of 2010. Finally, the main experiment for Breza et al. (2015) was rolled out during the summer and fall of 2010. Given the contained nature of each of these experiments, it is unlikely that they should interact in any way with the much more involved savings intervention. Nevertheless, we check for treatment balance and whether controls for past participation alter our results in any way.

Table M.1 also shows treatment balance by prior participation in the laboratory sessions. 29 of the study villages had participated in at least one of these three lab experimental sessions, with Endogenous selection villages exhibiting lower rates of prior participation in Panel A. Panel B further explores treatment balance at the saver level. 23% of households that enrolled in the savings study as savers had a household member who participated in at least one of the prior lab experiments. However, savers receiving a monitor in Random assignment villages are no more or less likely to have participated in the past. We do, however, detect treatment imbalance in receiving a monitor in Endogenous assignment villages.

We next check the robustness of our main results by including controls for prior household participation. Tables M.2, M.3, and M.4 present versions of Tables 2, 3, and 8, respectively, adding indicators for past participation by a household member in one, two, or three laboratory sessions. All of the key coefficients are very close in magnitude and of similar statistical significance to the main specifications in the body of paper. This gives us confidence that the prior laboratory experiments did not interact in a meaningful way with the savings intervention.

Table M.1. Treatments and Prior Experiments: Balance

Panel A: Village-level Balance

	Treatment	
	(1) Mean of Non-Monitored Savers	(2) Diff. Random vs. No Monitor
Microfinance Entered Village	0.633	-0.0667
	(0.0895)	(0.128)
Any Prior Lab Experiment	0.500	-0.200
	(0.0928)	(0.126)
Number Prior Lab Experiment	1.167	-0.367
	(0.204)	(0.286)

Panel B: Household-level Balance

	Treatment		
	(1) Mean of Non-Monitored Savers	(2) Diff. Random vs. No Monitor	(3) Diff. Endogenous vs. No Monitor
Any Prior Lab Experiment Participation			
No Fixed Effects	0.270	-0.013	-0.058
	(0.035)	(0.037)	(0.033)
Any Prior Lab Experiment Participation			
Village Fixed Effects	0.268	0.006	-0.073
	(0.021)	(0.042)	(0.032)
Number Prior Lab Experiment Participation			
No Fixed Effects	0.313	-0.008	-0.062
	(0.044)	(0.045)	(0.045)
Number Prior Lab Experiment Participation			
Village Fixed Effects	0.313	0.015	-0.090
	(0.026)	(0.049)	(0.046)

Notes: Table shows baseline balance by exposure to prior experiments. In each panel, the first column shows means and standard deviations in parentheses for the subgroup. Treatment differences are taken from regressions of each dependent variable on treatment indicators, with standard errors in parentheses. Panel A shows village-level characteristics, while Panel B shows saver-level characteristics. The observation count for Panel A is 60 villages and for Panel B is 1307 savers.

Table M.2. Treatments and Prior Experiments: Receipt of a Random Monitor

VARIABLES	(1) Log Total Savings	(2) Log Total Savings	(3) Log Total Savings	(4) Log Total Savings
Monitor Treatment: Random Assignment	0.347	0.275	0.344	0.272
	(0.154)	(0.162)	(0.156)	(0.165)
	[0.085, 0.610]	[-0.000, 0.551]	[0.079, 0.610]	[-0.009, 0.552]
Observations	549	549	549	549
Prior Experience Controls	Linear Controls	Linear Controls	Fixed Effects	Fixed Effects
Other Fixed Effects	None	Village	None	Village
Other Controls	None	Saver	None	Saver
Non-monitored mean	7.670	7.670	7.670	7.670

Notes: Table shows the effects of receiving a randomly allocated monitor on log total savings in the 30 random assignment villages. Total savings is the amount saved across all savings vehicles – the target account and any other account, both formal and informal including money held "under the mattress." Sample constrained to individuals who answered our questionnaire. Saver controls include the following saver characteristics: savings goal, saver centrality, age, marital status, number of children, preference for bank or post office account, whether the individual has a bank or post office account at baseline, caste, elite status, number of rooms in the home and type of electrical connection. Table also controls for prior participation in previous laboratory sessions. Columns 1 and 2 use linear controls in the number of prior sessions, while columns Standard errors (clustered at the village level) are reported in parentheses. 90% Confidence intervals are reported in brackets.

Table M.3. Treatments and Prior Experiments: Network Position of Random Monitor

VARIABLES	(1) Log Total Savings	(2) Log Total Savings	(3) Log Total Savings	(4) Log Total Savings
Monitor Centrality	0.176		0.131	
	(0.073)		(0.073)	
	[0.051, 0.300]		[0.008, 0.255]	
Saver-Monitor Proximity		1.041	0.875	
		(0.355)	(0.338)	
		[0.437, 1.645]	[0.301, 1.449]	
Model-Based Regressor (q_{ij})				0.206
				(0.120)
				[0.002, 0.411]
Observations	426	426	426	426
Prior Experience Controls	Fixed Effects	Fixed Effects	Fixed Effects	Fixed Effects
Other Fixed Effects	Village	Village	Village	Village
Other Controls	Saver and Monitor	Saver and Monitor	Saver and Monitor	Saver and Monitor

Notes: Table shows impacts on log total savings by monitor network position. Total savings is the amount saved across all savings vehicles – the target account and any other account, both formal and informal including money held "under the mattress." Sample constrained to savers who received a monitor in the 30 random-assignment villages, who answered our questionnaire. The variable "Model-Based Regressor" is defined as q_{ij} in the framework. Saver and Monitor controls include savings goal and saver centrality, along with the following variables for each monitor and saver: age, marital status, number of children, preference for bank or post office account (saver only), whether the individual has a bank or post office account at baseline, caste, elite status, number of rooms in the home and type of electrical connection. Saver and Monitor controls additionally include the geographical distance between their homes. Table also controls for prior participation in previous laboratory sessions. All columns include a saturated set of prior participation indicators. Standard errors (clustered at the village level) are reported in parentheses. 90% Confidence intervals are reported in brackets.

Table M.4. Treatments and Prior Experiments: Effects of Random and Endogenous Monitors

VARIABLES	(1) Log Total Savings	(2) Log Total Savings
Random Assignment Village × Monitor	0.276	0.285
	(0.152)	(0.148)
	[0.022, 0.531]	[0.038, 0.533]
Endogenous Assignment Village × Monitor	-0.125	-0.100
	(0.158)	(0.142)
	[-0.389, 0.139]	[-0.338, 0.138]
Endogenous Assignment Village		0.373
		(0.206)
		[0.029, 0.717]
Observations	1,061	1,061
Prior Experience Controls	Fixed Effects	Fixed Effects
Other Fixed Effects	Village	None
Other Controls	Saver	Saver

Notes: Table reports effects of receiving a monitor in random versus endogenous allocation villages. Total savings is the amount saved across all savings vehicles – the target account and any other account – by the saver. Sample includes savers who responded to our questionnaire. Prior experience controls are dummies for the number of previous experiments participated in. Other controls include the following saver characteristics: savings goal, saver centrality, age, marital status, number of children, preference for bank or post office account, whether the individual has a bank or post office account at baseline, caste, elite status, number of rooms in the home and type of electrical connection. Column 1 includes village fixed effects, while column 2 does not. Standard errors (clustered at the village level) are reported in parentheses. 90% Confidence intervals are reported in brackets.

Appendix N. Main Results, With Goal Trimming

Table N.1. Effect of Random Monitors on Savings: With Goal Trimming

VARIABLES	(1) Log Total Savings	(2) Log Total Savings	(3) Log Total Savings
Monitor Treatment: Random Assignment	0.370	0.284	0.353
	(0.146)	(0.162)	(0.138)
	[0.123, 0.618]	[0.009, 0.558]	[0.118, 0.587]
Observations	544	544	544
Fixed effects	None	Village	
Controls	None	Saver	Double-Post LASSO
Non-monitored mean	7.647	7.647	7.647

Notes: Table shows the effects of receiving a randomly allocated monitor on log total savings in the 30 random assignment villages. Total savings is the amount saved across all savings vehicles – the target account and any other account, both formal and informal including money held "under the mattress." Sample constrained to individuals who answered our questionnaire and whose saving goals were not in the top 1%. Saver controls include the following saver characteristics: savings goal, saver centrality, age, marital status, number of children, preference for bank or post office account, whether the individual has a bank or post office account at baseline, caste, elite status, number of rooms in the home and type of electrical connection. Standard errors (clustered at the village level) are reported in parentheses. 90% Confidence intervals are reported in brackets.

Table N.2. Total Savings by Network Position of Random Monitor: With Goal Trimming

VARIABLES	(1) Log Total Savings	(2) Log Total Savings	(3) Log Total Savings	(4) Log Total Savings
Monitor Centrality	0.178		0.134	
	(0.074)		(0.073)	
	[0.053, 0.303]		[0.010, 0.258]	
Saver-Monitor Proximity		1.032	0.865	
		(0.352)	(0.334)	
		[0.435, 1.630]	[0.298, 1.432]	
Model-Based Regressor (q_{ij})				0.217
				(0.118)
				[0.017, 0.417]
Observations	424	424	424	424
Fixed effects	Village	Village	Village	Village
Controls	Saver and Monitor	Saver and Monitor	Saver and Monitor	Saver and Monitor

Notes: Table shows impacts on log total savings by monitor network position. Total savings is the amount saved across all savings vehicles. Sample constrained to savers who received a monitor in the 30 random-assignment villages and who answered our questionnaire and whose saving goals were not in top 1%. The variable "Model-Based Regressor" is defined as q_{ij} in the framework. Saver and Monitor controls include savings goal and saver centrality, along with the following variables for each monitor and saver: age, marital status, number of children, preference for bank or post office account (saver only), whether the individual has a bank or post office account at baseline, caste, elite status, number of rooms in the home and type of electrical connection. Saver and Monitor controls additionally include the geographical distance between their homes. Standard errors (clustered at the village level) are reported in parentheses. 90% Confidence intervals are reported in brackets.

Appendix O. Discussion of Implementation Costs

One relevant policy consideration is the cost of implementing and scaling a peer monitoring product. Our specific treatments were implemented with research goals in mind and were never meant to be profitable or scalable. However, we do think that there are many opportunities for financial institutions to reduce the costs of product delivery. One of our largest costs was personnel. In order for the research team to have more control over the implementation, we chose to send individuals to each village on a bi-weekly basis to meet the savers, physically verify the passbooks, and pass the relevant information on to the monitors. Many financial institutions in India already use the Business Correspondent (BC) model, in which agents of the bank travel to villages to provide direct in-home customer service. This includes account opening procedures and deposit-taking. One could easily imagine a small tweak to this model, where the BC could intermediate information to others in the village after his pre-specified appointments. Further, banks could use technologies such as SMS to implement a peer monitoring scheme.

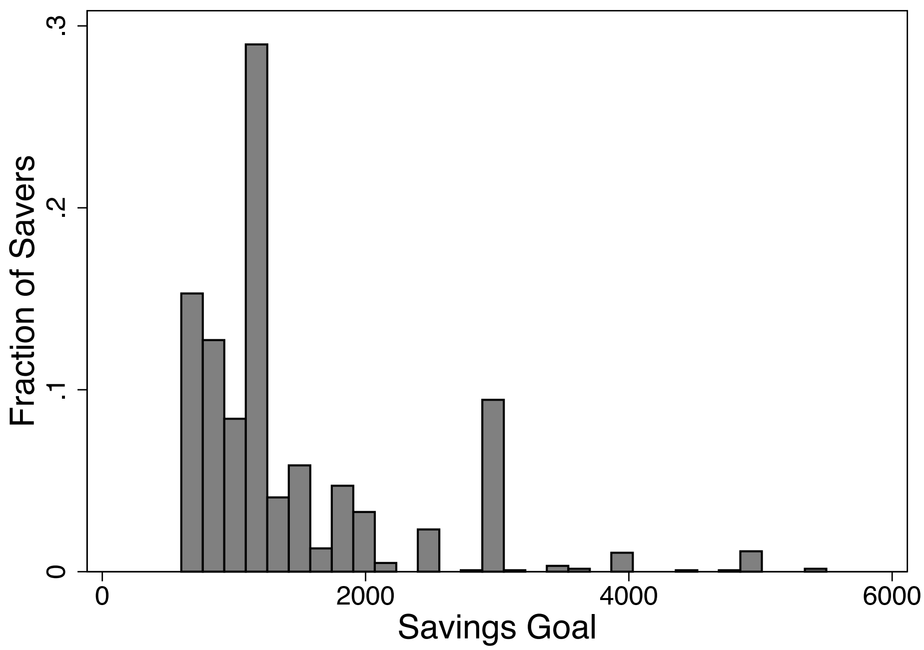
The other main cost associated with our intervention was the incentive given to monitors. First, as discussed previously, we think that the incentives had negligible effects on savings outcomes. Second, we certainly did not attempt to "optimize" the size of these incentives.[48] Nevertheless in the endogenous monitor case, the aggregate monitor incentives paid to participants correspond to a 6% semi-annual interest rate on all additional savings that were caused by our interventions, which – while not cheap – is not outlandish.[49] Experimenting with the size of the incentives would likely yield significant cost reductions.

[48] In fact, we believe that the optimal incentive would be close to, if not equal to, zero.

[49] To reach this 6% value, we first calculate the aggregate payments that we made to monitors in the Endogenous villages. We then calculate the excess savings across all savings vehicles that were caused by our treatments. We include both the direct effects of receiving a monitor on savings and also the spillovers onto non-monitored savers.

Appendix P. Baseline Savings Goals

Figure P.1. Histogram of Baseline Savings Goals



Notes: The figure shows the distribution of the baseline savings goals. We clip the top 5% tail of the distribution to make the figure more readable.

Figure P.1 presents the histogram of savings goals, censoring the top 5%.[50] There are a few large outliers (maximum goal Rs. 26,000), so the mean of Rs. 1838 shrinks to Rs. 1650 when we trim 1% outliers. In all specifications of our key results we drop the top 1% of savings goal observations.

[50]Note that the minimum goal is Rs. 600, the lower bound of allowed goals for participants.

Appendix Q. Negligibility of Terms

Our main model in Appendix A has two somewhat unattractive features, introduced for parsimony. First, it allows agents to obtain utility from meeting themselves and receiving payoffs from themselves. Second, when savers i meet their monitors j, we assume that information is transferred from the monitor to herself with probability $p_{jj} \leq 1$. Thus, the monitor may forget about the saver's performance. Here, we show that an alternate formulation that does not permit self-meetings and that allows monitors to have perfect recall about the saver's performance yields conclusions that, to first order, identically match those in Section A. The reason we prefer writing the core model in the style of Appendix A is simply for parsimony.

Q.1. **Utility.** Here we propose a modified utility function $U^*(s_i)$ such that savers i do not receive utility from meeting themselves, and when i and j meet, monitor j recalls the performance of saver i with probability 1. The expected payoff of saver i with monitor j for choice s_i is given by

$$U^*(s_i) := \sum_{k\neq i,j} p_{ik}p_{jk}\mathrm{E}_k\left[A_{\theta_i}|s_i, r_k = 1\right] + p_{ij}\mathrm{E}_j\left[A_{\theta_i}|s_i, r_j = 1\right] + \sum_{k\neq i,j}\left(1 - p_{jk}\right)p_{ik}\mathrm{E}_k\left[A_{\theta_i}|s_i, r_k = 0\right] - s_i c_{\theta_i}.$$

Note that

$$\begin{aligned} U(s_i) - U^*(s_i) &= p_{ii}p_{ji}\mathrm{E}_i\left[A_{\theta_i}|s_i, r_i = 1\right] \\ &\quad + \left(1 - p_{ji}\right)p_{ii}\mathrm{E}_i\left[A_{\theta_i}|s_i, r_k = 0\right] \\ &\quad - \left(1 - p_{jj}\right)p_{ij}\left(\mathrm{E}_j\left[A_{\theta_i}|s_i, r_j = 1\right] - \mathrm{E}_j\left[A_{\theta_i}|s_i, r_j = 0\right]\right) \end{aligned}$$

Q.2. **Defining q^*_{ij}.** Given this alternate utility formulation, $U^*(s_i)$, we can rewrite Lemma A.1 and Proposition A.2 using q^*_{ij} rather than q_{ij}, where

$$q^*_{ij} = \sum_{k\neq i,j} p_{ik}\cdot p_{jk} + p_{ij}.$$

This captures the cases where the saver i interacts with third party k in the second phase and third party k has been informed by the monitor j in the first phase. The last term reflects the fact that the monitor knows the saver's performance already (by definition), so all we need is to calculate the probability that the saver i interacts with the monitor k in the second phase.

But note that

$$\begin{aligned} q^*_{ij} &= \sum_k p_{ik}\cdot p_{jk} - p_{ii}p_{ji} - p_{ij}\left(p_{jj} - 1\right) \\ &= q_{ij} - p_{ii}p_{ji} - p_{ij}\left(p_{jj} - 1\right). \end{aligned}$$

Q.3. **The Behavior of** $E_{ij}[\frac{q_{ij}-q^*_{ij}}{q_{ij}}]$**.** Our goal is to show that, with high probability, the average value (across ij) of the remainder term $q_{ij} - q^{\star}_{ij}$ is negligible relative to q_{ij} in large graphs.[51]

We will consider a sequence of random graphs from the Chung-Lu model (Chung and Lu, 2001): More precisely, for each n, fix a tuple $(w^n_1, \ldots, w^n_n) \in \mathbb{R}^n_{\geq 0}$, satisfying $\max_i [w^n_i]^2 \leq \sum_{k=1}^n w^n_k$. Construct a sequence of (random) adjacency matrices $\mathbf{A}^{(n)}$ as follows: the probability of an edge ij for $i \neq j$ is:

$$\mathrm{P}\left(A^n_{ij} = 1\right) = \frac{w^n_i w^n_j}{\sum_k w^n_k},$$

and all these realizations are independent. There are no self-loops, so $A^{(n)}_{jj} = 0$ throughout. We further assume that there is a constant $\bar{d} > 1$ such that $1 \leq w^n_i < \bar{d}$ for every $i, n \in \mathbb{N}$.[52] This makes the maximal expected degree uniformly bounded over the sequence. This is a useful model to consider as it is a sequence of sparse graphs, which is consistent with real-world observations, and allows for flexible degree sequences in the space of sparse graphs.

Recall from Section A that $\mathbf{M}$ and $\mathbf{P}$ are defined as functions of $\mathbf{A}$ as follows: $\mathbf{M} := \sum_{t=1}^T (\theta \mathbf{A})^t$ for $\theta \in (0,1)$, $\mathbf{P} := \beta \mathbf{M}$. Recall also that $\beta = \beta_n$ is a sequence of fractions that tends to zero as $n \to \infty$.

Proposition Q.1. *Consider the sequence of random graph models indexed by n as above, under the above assumptions. Let $\mathbf{A} = \mathbf{A}^{(n)}$ denote a draw from the nth such model.*[53] *Let T (in the definition of $\mathbf{M}$) be constant in n. Finally assume that $\frac{1}{n\beta_n} = o(1)$. Then*

$$\mathrm{E}_{ij}\left[\frac{p_{ii}p_{ji} + p_{ij}(p_{jj} - 1)}{p_{ii}p_{ji} + p_{ij}p_{jj} + \sum_{k \neq i,j} p_{ik}p_{jk}}\right] = o(1)$$

where $\mathrm{E}_{ij}[\cdot]$ is the expectation over A of the empirical mean over i, j.

Proof. It is sufficient to show that there is some function $f : \mathbb{N} \to \mathbb{N}$ such that for all $C > 0$ and any $\delta > 0$, there exists some $\bar{n} \in \mathbb{N}$ such that, for all $n \geq \bar{n}$, the share of (i, j) satisfying the following two conditions is at least $1 - \delta$:

(1) $\sum_{k \neq i,j} p_{ik}p_{jk} > Cf(n)$ and
(2) $|p_{ii}p_{ji} + p_{ij}(p_{jj} - 1)| \leq f(n)$.

To show (2), it is sufficient to bound the highest-order term, which will be p_{ij}. Let $d_n := \mathrm{E}\left[\frac{1}{n}\sum_{i,j} A_{i,j}\right] = \frac{\sum_k w^n_k}{n}$. Under the above assumptions $1 \leq d_n \leq \bar{d}$.

Then we have the following sequence of statements

---

[51] We thank Ben Golub for helping us formalize the argument below.

[52] In this model, the weight w^n_i is the expected degree of node i in $\mathbf{A}^{(n)}$ since $\mathrm{E}\left[\sum_j A^n_{ij}\right] = \frac{\sum_j w^n_i w^n_j}{\sum_k w^n_k} = w^n_i$.

[53] We will sometimes suppress indices for n in matrices and in β_n to avoid clutter.

$$\begin{aligned}
\mathrm{E}_{ij}\left[p_{ij}\right] = \frac{\beta}{n^2}\sum_{i,j}\mathrm{E}\left[M_{ij}\right] &= \frac{\beta}{n^2}\mathrm{E}\left\{\sum_{i,j}\sum_{t=1}^{T}\left[\theta^t\mathbf{A}^t\right]_{ij}\right\} \\
&= \frac{\beta}{n^2}\sum_{t=1}^{T}\theta^t\mathrm{E}\left\{\sum_{i,j}\left[\mathbf{A}^t\right]_{ij}\right\} \\
&\leq K'\cdot\frac{\beta}{n}\sum_{t=1}^{T}\theta^t\cdot d^t \text{ wpa1} && \text{(a), see below} \\
&\leq K''\cdot\frac{\beta}{n}\cdot\bar{d}^T && \text{(b), see below} \\
&\leq K\cdot\frac{\beta}{n} && \text{(c), see below}
\end{aligned}$$

for constants $K, K', K'' > 0$ which do not depend on n.

Here (a) follows from the fact that each of the sums $\sum_{i,j}\left[\mathbf{A}^t\right]_{ij}$ for $t \leq T$ are random variables with finite variance and a mean proportional to d^t. Then by Markov's inequality

$$\mathrm{P}\left\{\left|\frac{1}{n}\sum_{i,j}\left\{\left[\mathbf{A}^t\right]_{ij} - \text{const}\cdot d^t\right\}\right| > \eta\right\} \leq \frac{\mathrm{E}\left|\sum_{i,j}\left\{\left[\mathbf{A}^t\right]_{ij} - \text{const}\cdot d^t\right\}\right|}{n\cdot\eta}.$$

The statement (b) follows from the fact that $d^t \leq \bar{d}^t \leq \bar{d}^T$, and (c) follows by absorbing that into the constant.

For (1), it is sufficient to bound $\mathrm{E}_{i,j}\left[\sum_{k\neq i,j} p_{ik}p_{jk}\right]$, which we begin doing through the following set of algebraic transformations.

$$\begin{aligned}
\mathrm{E}_{i,j}\left[\sum_{k\neq i,j} p_{ik}p_{jk}\right] &= \frac{1}{n^2}\sum_{i,j}\sum_{k\neq i,j}\beta^2\mathrm{E}\left\{\left[\sum_{t=1}^{T}\left[\theta^t A^t\right]_{ik}\right]\left[\sum_{t'=1}^{T}\left[\theta^{t'} A^{t'}\right]_{jk}\right]\right\} \\
&= \frac{\beta^2}{n^2}\sum_{s=2}^{2T}\theta^s\sum_{t=1}^{\min\{s,T\}}\mathrm{E}\left\{\sum_{i,j}\sum_{k\neq i,j}\left[A^t\right]_{ik}\left[A^{s-t}\right]_{jk}\right\}.
\end{aligned}$$

Our goal is to then bound the above expression by an expression depending on β, n, $\bar{d}$, and constants. We will do this by treating $\left[A^t\right]_{ik}$ and $\left[A^{s-t}\right]_{jk}$ as essentially independent, so that we can split up (modulo negligible error) the expectation of the product as the product of expectations. Consider a walk W from i to k. Call the random set of nodes appearing in this walk V_W. Conditional on there a walk from j to k, also of length T, because maximum degree is bounded, the probability of this walk including any of the nodes in V_W is $O(1/n)$. Applying the union bound (across nodes and then across walks) shows that the probability of any walk from i to k sharing any nodes with a walk from j to k is $O(1/n)$.[54] Here it is important that there are, in expectation, finitely many walks.

[54]See Cooper et al. (2009).

Thus, we have[55]

$$\frac{\beta^2}{n^2}\sum_{s=2}^{2T}\theta^s\sum_{t=1}^{\min\{s,T\}}\mathrm{E}\left\{\sum_{i,j}\sum_{k\neq i,j}\left[A^t\right]_{ik}\left[A^{s-t}\right]_{jk}\right\}$$

$$=\frac{\beta^2}{n^2}\sum_{s=2}^{2T}\theta^s\sum_{t=1}^{\min\{s,T\}}\left\{\sum_{i,j}\sum_{k\neq i,j}\mathrm{E}\left[A^t\right]_{ik}\mathrm{E}\left[A^{s-t}\right]_{jk}\right\}\qquad \text{wpa 1}$$

$$=\Theta\left(\beta^2\sum_{s=2}^{2T}\theta^s\sum_{t=1}^{\min\{s,T\}}(n-1)\,d^t\,(n-1)\,d^{s-t}\right)$$

$$=K'''\frac{\beta^2(n-1)^2}{n^2}\sum_{s=2}^{2T}\theta^s d^s\sum_{t=1}^{\min\{s,T\}}\cdot 1\qquad \text{(d), see below}$$

for constant $K'''>0$,which does not depend on n.

Now notice that

$$\frac{\beta^2(n-1)^2}{n^2}\cdot\underbrace{\theta^{2T}d}_{\text{constant in } n}\leq\frac{\beta^2(n-1)^2}{n^2}\sum_{s=2}^{2T}\theta^s d^s$$

$$\leq\frac{\beta^2(n-1)^2}{n^2}\sum_{s=2}^{2T}\theta^s d^s\sum_{t=1}^{\min\{s,T\}}\cdot 1$$

$$\leq\frac{\beta^2(n-1)^2}{n^2}\cdot\underbrace{2T\cdot\bar{d}^{2T}}_{\text{constant in } n}$$

so the expression (d) is $\Theta(\beta_n^2)$ in n.

Now

$$\frac{\beta/n}{\beta^2}\to 0$$

by our assumption that our assumption that $\frac{1}{n\beta_n}=o(1)$. Take the sequence $f(n)=\frac{\beta_n}{n}$; our argument then shows that (1) and (2) hold for this sequence since (c) provides an upper bound proportional to $f(n)=\beta_n/n$ for $\mathrm{E}_{ij}[p_{ij}]$ and (d) provides a lower bound proportional to β_n^2 for $\mathrm{E}_{i,j}\left[\sum_{k\neq i,j}p_{ik}p_{jk}\right]$. Under our assumption then, β_n^2 asymptotically dominates $f(n)=\beta_n/n$, then the result follows. □

[55]Recall in Bachmann-Landau notation, we say $g(n)=\Theta(h(n))$ if $\exists k_1>0,\exists k_2>0,\exists n_0$, s.t. $\forall n>n_0$,

$$k_1\cdot h(n)\leq g(n)\leq k_2\cdot h(n).$$

This means g is bounded both above and below by h asymptotically.