Online Appendix:

Adoption with Social Learning and Network Externalities*

Marcel Fafchamps — Stanford University

Måns Söderbom — University of Gothenburg

Monique van den Boogaart — Uber, San Fransisco

December 2021

*This is the Online Appendix for the paper "Adoption with Social Learning and Network Externalities," by Marcel Fafchamps, Måns Söderbom and Monique van den Boogaart. This paper is the "main text" referred to in the present Appendix.

Appendix A. Conceptual framework

Appendix A provides the theoretical framework that motivates our empirical analysis. The focus of our attention is adoption, that is, the first usage of a new product or service by someone who has not used it before. We are interested in how social networks influence adoption. We propose a model of social learning inspired, among others, by the work of Young (2009), except that our model explicitly distinguishes between learning about the existence of a product and learning about its attributes.[1] Since the purpose of the model is solely to provide a theoretical basis for our identification strategy, we ignore the equilibrium properties of this learning process and abstract from the topological properties of the social network (see, e.g., Jackson and Yariv 2005, Jackson 2008, Kreindler and Young 2014 and Arieli et al. 2020 for details).

In the first part of the presentation, we ignore the possibility of network externalities in usage and we focus exclusively on social learning. Since network externalities and other sources of correlation in usage are regarded as confounds in our testing strategy, they are only discussed briefly at the end.

Formally, let usage of a product $y_{it} = \{0, 1\}$ be a dichotomous variable equal to 1 if individual i uses the product at time t, and 0 otherwise. We think of time as a sequence of time intervals, e.g., a week. Adoption describes the first time at which $y_{it} > 0$ for individual i. Let t_i denote the time at which individual i becomes 'at risk' of adopting the product.[2] Further let T_i denote the time at which individual i first uses the product. Finally, let T denote the last data period for which we have information. By definition, $T_i > T$ for an individual who, by time T, has not yet used the product. As we will argue below, usage after adoption provides useful information

[1] Young (2009) also considers diffusion by contagion or conformism. They are ignored here since we do not regard them as relevant in our context. To the extent that they do exist in our study, however, they would presumably apply to usage as well and, as such, would be subsumed in the network extenalities we discuss below.

[2] This can be the time at which the new product is introduced, or the time at which i acquires a device for which product is useful.

as well. Usage y_{it} can therefore be divided into two vectors or periods: the time until first usage $\{y_{it_i}, ...y_{iT_i}\}$; and usage after that $\{y_{iT_i+1}, ...y_{iT}\}$. By construction, $\{y_{it_i}, ...y_{iT_i}\}$ is either a sequence of 0's ending with a single 1, or a string of 0's (for someone who never adopts). The length of each of the two i vectors varies across individuals.

We are interested in identifying predictors of y_{it} that depend on the adoption and usage behavior of the social neighbors of i. To do so effectively, we present a few simple concepts before articulating our testing strategy. We first discuss social learning, before introducing network externalities. We assume throughout that the researcher has information about y_{it}.

Social learning about product existence

There is much to learn from simple models of social learning. Let us first focus on information about the existence of the product. We then turn to information about the qualities of the product. We end with a short discussion of experimentation, which is adoption purely for the purpose of eliciting information about product quality. The focus of this section is to use simple models to develop intuition about social learning that we can then take to the data. The building blocks of the model are inspired by Young (2009).

Learning about the existence of the new product closely resembles a contagion process (e.g., Bass 1969). Without information about the existence of the product, the agent simply cannot adopt. Hence having been exposed to information about the product is a necessary condition for adoption. This information can come from two sources: (1) information received from various sources outside the social network (e.g., ads on billboard, radio, TV, junk mail, or newspaper); and (2) information received from the social network (e.g., friends, relatives, co-workers).

Let θ_{vt} denote the probability of receiving information from outside the social network in location v at time t. We take this probability as given and we do not seek to model its de-

terminants. But we think of it as having a strong local component, capturing the local nature of advertisement coverage. A simple model for the probability of receiving information from a social source at time t can be formulated as:

$$\Pr(i \text{ receives information from network at } t+1) = 1-(1-q)^{\Delta A_{it}}$$

where ΔA_{it} is the number of neighbors of i who have started using the product in period t – and thus have become aware of its existence and can relay this information to i, something each of them does with probability q. We assume that the researcher observes ΔA_{it}, or a close proxy. The cumulative probability that i has received information about the existence of the product is thus an increasing and concave function of the cumulative number of i's neighbors who have adopted at t – and thus could have passed information about the product to i with probability q during that time period.

Let us now combine the two sources of information. If we assume independence between θ_{vt} and the signal received from each neighbor, the probability of *not* being informed within period t is $(1-\theta_{vt})(1-q)^{\Delta A_{it}}$. Now let us assume that, once i is informed that the product exists, i adopts with probability p_i. This is the probability of usage in any given period, conditional on knowing about the product. For some individuals this probability is low; for others it is high.

Over time the likelihood of having heard of the product increases. Formally, the probability of *not* having heard of the product between time t_i and t is:

$$\Pr = \prod_{s=t_i}^{t}(1-\theta_{vs})(1-q)^{\Delta A_{is}} = (1-q)^{A_{it}}\prod_{s=t_i}^{t}(1-\theta_{vs})$$

where $A_{it} \equiv \sum_{s=t_i}^{s=t}\Delta A_{is}$ is the cumulative number of adopting neighbors between t_i and t, and t_i is the time at which i starts being at risk of being exposed to information about the product's

existence. If θ_{vt} is constant over time for location v, the formula simplifies to:

$$\Pr = (1-q)^{A_{it}}(1-\theta_v)^{S_{it}}$$

where $S_{it} \equiv t - t_i$ is the time elapsed between t_i and t.

The probability that agent i adopts the product at time t is the probability that he has been informed times p_i:

$$\Pr(y_{it+1} = 1 | \{y_{it_i}, ..., y_{it}\} = \{0, ..., 0\}) = [1 - (1-q)^{A_{it}}(1-\theta_v)^{S_{it}}]p_i \tag{A1}$$

Adoption can take place even for someone who has no social neighbors, or whose neighbors have not adopted. The model predicts that the likelihood of adoption increases in a systematic fashion over time, without or without adopting neighbors. This is a mechanical effect: as time passes, the agent has more and more chances of hearing about the product. The probability of first adoption increases with time since inception S_{it} and with A_{it}, although in both cases the effect is concave: the derivative of the probability of adoption with respect to. S_{it} and A_{it} falls with S_{it} and with A_{it}. This is because having heard about the product once is enough to know of its existence.

Once the product has been used once, i may continue using it with a certain probability. But if the only source of network effects is social learning about the existence of the product, the probability of usage after first adoption is no longer a function of the number of adopting neighbors. Formally we have:

$$\Pr(y_{it+1} = 1 | y_{is} = 1 \text{ for some } s < t) \ = p_i + \varepsilon_{it+1} \tag{A2}$$

Thus once i has learned about the existence of the product, the data generating process shifts from (A1) to (A2). An identical prediction is made if the researcher observes a signal M_{it} that is equal to 1 when individual i has unambiguously been made aware of the existence of the new product, and 0 otherwise:

$$\Pr(y_{it+1} = 1 | M_{is} = 1 \text{ for some } s < t) \ = p_i + \varepsilon_{it+1} \tag{A3}$$

To recap, when network neighbors circulate information about product existence and nothing more, the probability of adoption increases in the number of adopting neighbors, but at a decreasing rate. After first adoption or after becoming aware of the product, subsequent usage does not depend on the number of adopting neighbors.

Social learning about product quality

We get different predictions if social learning is about product quality. In this case, the decision to adopt at time t depends not on the probability of receiving a signal within a given time interval, but rather on the cumulative information about the product received up to time t (e.g., Bala and Goyal 1998, Jackson 2008).

To keep the same notation, let θ_{vt} now denote the probability that individual i receives an independent signal about the quality of the product at time t. This probability can vary over time t and across locations v. To keep things simple, let us assume that this signal takes only two values, 0 and 1, i.e., a bad signal or a good signal. Let μ denote the true probability that the product performs: a high μ good always performs well, while a low μ good often performs poorly. Individuals differ in how much they value unobserved quality μ – more about this later. We assume that the posterior belief h_{it} of individual i at time t is simply the sample estimate

of the unknown Bernoulli parameter μ based on the information available to i at time t.[3] Let N_{it} be the number of signals received by i at up to t and let N_{it}^1 be the number of signals with value 1, i.e., the number of good signals. We have:

$$h_{it} = \frac{N_{it}^1}{N_{it}} \tag{A4}$$

The variance of this belief is approximately given by:

$$v_{it}^2 = \frac{1}{N_{it}} h_{it}(1 - h_{it}) \tag{A5}$$

As sample size increases, h_{it} tends to μ and v_{it}^2 tends to 0.[4]

Since we do not observe what signal people observe, we never know what N_{it}^1 is. But we can write:

$$h_{it} = \mu + e_{it} \text{ with } e_{it} \tilde{} (0, \mu(1-\mu)/N_{it})$$

In other words, the information people have is, on average, unbiased and the variance of their beliefs shrinks over time.

If we allow agents to hold a prior belief h_{i0}, this belief can be regarded as coming from a sample of observations N_{i0} that we do not observe. The point estimate of this belief marks how biased the prior belief is, and the size of the sample determines how confident the agent is in his

[3] This is simplified Bayesian approach – see Mood, Graybill and Boes (1974) p. 342 for the correct Bayesian estimator of a Bernouilli parameter. But this simple approach suffices for our purpose.

[4] The above formula for the variance is obtained by combining Mood et al. (1974) p. 236 with p. 89.

prior belief. This can be formalized as follows:

$$\begin{aligned} h_{i0} &= \frac{N^1_{i0}}{N_{i0}} \\ h^b_{it} &= \frac{N^1_{i0}+N^1_{it}}{N_{i0}+N_{it}} = h_{i0}\frac{N_{i0}}{N_{i0}+N_{it}} + h_{it}\frac{N_{it}}{N_{i0}+N_{it}} \\ v^2_{it} &= \frac{1}{N_{i0}+N_{it}} h^b_{it}(1-h^b_{it}) \end{aligned}$$

where h^b_{it} now denotes the posterior belief of agent i at t. We do not observe h_{i0} and N_{i0}. If we let the number of signals received be denoted n_{it}, beliefs can be written as following a model of the form:

$$\begin{aligned} q^b_{it} &= \alpha\frac{\gamma}{\gamma+n_{it}} + \mu\frac{n_{it}}{\gamma+n_{it}} + e^b_{it} \text{ with } e^b_{it} \tilde{} (0, \sigma^2_{it}) \\ \sigma^2_{it} &= \frac{1}{\gamma+n_{it}}(\alpha\frac{\gamma}{\gamma+n_{it}} + \mu\frac{n_{it}}{\gamma+n_{it}})(1-\alpha\frac{\gamma}{\gamma+n_{it}} - \mu\frac{n_{it}}{\gamma+n_{it}}) \end{aligned}$$

As with uninformed priors, beliefs h^b_{it} tend to μ over time, but they show some persistence around initial priors.[5]

Having modelled learning, we now turn to adoption. We start without prior beliefs. We assume that individuals differ in the threshold value of μ that they require before adopting. At first glance, it seems that we could simply assume that people adopt if their estimate of μ is larger than some value τ_i with $0 < \tau_i < 1$. This decision rule, however, is too crude. It predicts that people adopt after a single good signal since, in that case, their posterior belief is $h_{i1} = 1 \geq \tau_i$ for any τ_i. This is clearly an unappealing decision rule because an estimate of μ based on a single observation is very imprecise. To capture this intuition in the simplest

[5] The variance σ^2_{it} is not monotonic over time, however. Intuition is as follows. Imagine the agent starts with a strong prior far from μ (a strong prior means N_{i0} is large). Initially σ^2_{it} is quite small because it is dominated by the strong prior. As more information is revealed, posterior beliefs are progressively pulled away from prior h_{i0} and σ^2_{it} increases. Eventually posterior beliefs settle on μ and the variance falls, dominated now by N_{it}.

possible way, we posit that the expected utility of adoption $E[U_{it}(y_{it} = 1)|\omega_{it}]$ can be written as a mean-variance form. We have:

$$y_{it+1} = 1 \text{ iff } h_{it} - Rv_{it}^2 \geq \tau_i$$

where R is a risk aversion parameter and τ_i is now a threshold value of expected utility. Since we do not observe h_{it} and v_{it}^2 directly, we replace them by formulas (A4) and (A5) above and we get:

$$\Pr(y_{it+1} = 1|\{y_{it_i}, ..., y_{it}\} = \{0, ..., 0\}) = \Pr\left((\mu - \tau_i) - R\frac{\mu(1-\mu)}{n_{it}} \geq -e_{it+1}\right) \tag{A6}$$

Equation (A6) shows that the probability of adoption increases with n_{it}. The intuition is straightforward: the variance term shrinks and vanishes at the limit, and this raises the expected utility of adoption for some people. Not everybody adopts, however, because μ is not higher than τ_i for everyone.

We can now generalize the above to the case where people hold prior beliefs. We now have:

$$\Pr(y_{it+1} = 1|\{y_{it_i}, ..., y_{it}\} = \{0, ..., 0\}) = \tag{A7}$$
$$\Pr\left(\alpha\frac{\gamma}{\gamma + n_{it}} + \mu\frac{n_{it}}{\gamma + n_{it}} + R\frac{1}{\gamma + n_{it}}(\alpha\frac{\gamma}{\gamma + n_{it}} + \mu\frac{n_{it}}{\gamma + n_{it}})(1 - \alpha\frac{\gamma}{\gamma + n_{it}} - \mu\frac{n_{it}}{\gamma + n_{it}}) \geq \tau_{it} - e_{it+1}^b\right)$$

To close the model, we need to stipulate the data generating process of n_{it}, the number of signals received. In practice, we do not observe n_{it} but, by analogy with the previous sub-section, we expect it to be an increasing function of time since inception S_{it} and of the number of adopting neighbors A_{it}. To show this formally, let us assume that in each period individual i receives

a signal from outside his network with a constant location-specific probability θ_v,[6] and with probability q individual i receive a signal from any newly adopting neighbor. The expected number of signals received at time t is a sum of two binomial processes. The average number of signals received outside the network up to time is given by a binomial process with parameter θ_v and S_{it}, and is simply $\theta_v S_{it}$. The average number of signals from the networks is qA_{it}. Thus we have:[7]

$$n_{it} = \theta_v S_{it} + qA_{it} + u_{it} \text{ with } u_{it} \tilde{} (0, v^2) \tag{A8}$$

Without prior beliefs, the probability of adoption can thus be written:

$$\Pr(y_{it+1} = 1 | \{y_{it_i}, ..., y_{it}\} = \{0, ..., 0\}) = \Pr\left((\mu - \tau_i) - R\frac{\mu(1-\mu)}{\theta_v S_{it} + qA_{it} + u_{it}} \geq -e_{it+1}\right) \tag{A9}$$

Equation (A9) shows that the probability of first adoption is monotonically increasing in S_{it} and A_{it}.

The probability of adoption with prior beliefs is similarly obtained by replacing n_{it} in equation (A7) by its value given by (A8). Our earlier observation remains valid: with strong prior beliefs, the variance term that multiplies R in equation (A7) can initially be quite small. If the prior belief h_{i0} is high and its variance v_{i0}^2 is small, individual i will adopt immediately. The social learning model therefore predict that individuals with strong optimistic priors adopt early. So doing, they receive information about the quality of the product, information that they may circulate among their social circle. If the information is sufficiently bad, i.e., if revealed quality is less than τ_i, early adopters will abandon the new product, and the information that diffuses

[6] To keep the algebra simple and derive the intuition clearly, we ignore here the possibility of a time-varying signal probability.

[7] Where, given our assumptions, v^2 can in principle be calculated from the variance formula for binomial distributions.

among the social network will discourage adoption by others. If the information is sufficiently good, its diffusion in the network will progressively raise posterior beliefs according to equation (A7) and adoption will spread among individuals with a sufficiently high valuation τ_i for the product. Because the accumulation of information eventually reduces the variance of posterior beliefs, adoption is an increasing function of the information received, and thus of the number of adopting neighbors.

What happens after an individual has adopted the product once? In the context of our empirical application, it is natural to assume that usage reveals a lot of relevant information about the product. To capture this idea in a stylized way, let us imagine that using the product once perfectly reveals the quality of the product. It follows that usage is now driven by τ_i; social learning no longer matters. Formally we have:

$$\Pr(y_{it+1} = 1 | y_{is} = 1 \text{ for some } s \leq t) \ = \Pr\left((\mu - \tau_i) \geq -e_{it+1}\right) \tag{A10}$$

which does not depend on time or adopting neighbors.

What happens if individual i is observed to receive an unambiguous signal revealing the existence of the product? In this case, this signal does not, by itself, dispel uncertainty about the quality of the product and thus should not eliminate the role of social learning in reducing uncertainty about the net benefit of adoption. In other words, adoption continues to follow equation (A7) after $M_{it} = 1$. This is different from what happens when social learning only affects knowledge about the existence of the product, and thus provides a way of identifying which type of social learning is present in the data.

To summarize, when social learning is purely about product quality, the likelihood of adoption is predicted to increase over time as the number of adopting neighbors rises, irrespective of whether the individual received a signal about product existence or not, that is, whether $M_{is} = 1$

or not. After first adoption, however, the role of social learning essentially disappears and the probability of continued usage is no longer a function of the number of adopting neighbors. In contrast, if social learning is solely about product existence, the data generating process switches to (A3) after $M_{is} = 1$. This makes it possible to test the two learning models against each other even in a reduced form. If social learning combines both elements, then we expect the coefficient of A_{it} to be significantly lower after $M_{is} = 1$, but to remain positive until first adoption.

Network externalities and strategic complementarities

Social learning can be seen as a network externality: individuals benefit from the information accumulated and shared by others. We have shown that social learning generates a correlation between neighbors' adoption and own adoption by individual i. There are many other network externalities that similarly induce strategic complementarities but do not involve learning. The main distinction between these other strategic complementarities and social learning is that social learning disappears after i has used the product at least once, while other strategic complementarities do not. This simple observation forms the basis of our identification strategy between social learning and other network externalities, as explained in the Testing Strategy Section of the main text.

References

Arieli, Itai, Yakov Babichenko, Ron Peretz, and H. Peyton Young (2020). "The Speed of Innovation Diffusion in Social Networks", *Econometrica*, 88(2): 569-94.

Bala, Venkatesh and Sanjeev Goyal (1998), "Learning from Neighbours", *Review of Economic Studies*, 65(3): 595-621.

Bass, Frank (1969). "A New Product Growth for Model Consumer Durables," *Management Science* 15(5), Theory Series, pp. 215-227.

Jackson, Matthew O. (2008). *Social and Economic Networks*, Princeton University Press, Princeton, 2008.

Jackson, Matthew O. and Leeat Yariv (2005) "Diffusion on Social Networks", *Economie Publique*, 16(1): 3-16.

Kreindler, Gabriel E. and H. Peyton Young (2014). "Rapid Innovation Diffusion in Social Networks", *PNAS*, 111(3): 10881-10888.

Mood, Alexander M., Franklin A. Graybill and Duane C. Boes (1974). *Introduction to the Theory of Statistics*, Third edition, McGraw-Hill, New York.

Young, H. Peyton (1999). "Diffusion in Social Networks," Papers 2, Brookings Institution - Working Papers.

Appendix B: Contemporaneous common shocks

Arguably the biggest threat to our identification strategy is common shocks. As explained in the main text, we deal with this issue in several ways. We first-difference our regressions to net out any highly persistent common shocks affecting the dependent variable y_{it} and our main regressors of interest; we include a large number of dummy variables controling for shocks shared across various geographical entities; and we use the correlation between Δy_{it} and ΔA_{it} *after* adoption as control for any added effect of social learning before adoption. In this section of the Appendix we discuss in detail how we address the possibility of remaining contemporaneous common shocks between y_{it} and A_{it}.

Common shocks and usage

We start by examining the effect of contemporaneous common shocks on usage. To keep the notation simple, we consider a case in which individual i has a single neighbor. Extending to multiple neighbors is straightforward. Let y^*_{it} be a latent variable reprensenting the benefit that user i expects to derive from ME2U at time t, and let y^*_{jt} be the expected benefit to user j who is the sole neighbor of i. Imagine that the benefit from usage is subject to a common contemporaneous shock c_t:

$$y^*_{it} = c_t + e_{it}$$

$$y^*_{jt} = c_t + e_{jt}$$

where c_t is an i.i.d. shock, common to i and j, with mean 0 and standard deviation σ_c and e_{it} is an i.i.d. idiosyncratic shock with mean 0 and standard deviation normalized to 1.[8] There

[8] This normalization of the standard deviation is without loss of generality since, in practice, we do not observe y^*_{it}.

are no social learning or network effects in this model; any correlation in adoption and/or usage between i and j would thus be the result of correlated shocks. Usage of ME2U is denoted by the dichotomous variable $y_{it} = \{0, 1\}$, defined by:

$$Pr(y_{it} = 1) = 1[y^*_{it} > \tau] \tag{B1}$$

where 1[.] is an indicator function and τ is the threshold value of y^*_{it} above which usage takes place.

It is immediately apparent that y^*_{it} and y^*_{jt} are positively correlated since they share the common shock c_t. This property also holds for $\Delta y^*_{it} \equiv y^*_{it} - y^*_{i,t-1}$ and Δy^*_{jt} since they share common term $c_t - c_{t-1}$. Lagging Δy^*_{jt} by one period, however, changes the nature of this correlation since:

$$\Delta y^*_{it} = c_t - c_{t-1} + e_{it} - e_{i,t-1} \tag{B2}$$

$$\Delta y^*_{jt-1} = c_{t-1} - c_{t-2} + e_{jt-1} - e_{j,t-2} \tag{B3}$$

The covariance between Δy^*_{it} and Δy^*_{jt-1} is now *negative* since the only shared term is c_{t-1} which appears with opposite signs in the two expressions. It follows that if we regress Δy^*_{it} on Δy^*_{jt-1}, we obtain a negative coefficient, the magnitude of which depends on σ_c. These properties are inherited by y_{it} and y_{jt}. Based on this, a *positive* correlation between Δy_{it} on Δy_{jt-1} cannot be accounted for by contemporaneous common shocks. Contemporaneous common shocks do not, however, induce any correlation between Δy^*_{it} and Δy^*_{jt-2} – or between Δy_{it} on Δy_{jt-2} – since, as shown in equation (B4), they do not include any common terms.

$$\Delta y^*_{jt-2} = c_{t-2} - c_{t-3} + e_{jt-2} - e_{j,t-3} \tag{B4}$$

This can be generalized to common shocks extending over two periods. Let's now assume that:

$$y^*_{it} = c_t + e_{it} + \gamma c_{t-1} \tag{B5}$$

$$y^*_{jt} = c_t + e_{jt} + \gamma c_{t-1} \tag{B6}$$

where $\gamma \leq 1$ captures the effect of the lagged common shock on current expected benefit from usage. Again we see that y^*_{it} and y^*_{jt} are positively correlated since they share the term $c_t + \gamma c_{t-1}$. Lagging Δy^*_{jt} by one period we now have:

$$\Delta y^*_{it} = c_t - (1-\gamma)c_{t-1} + e_{it} - e_{i,t-1} - \gamma c_{t-2}$$

$$\Delta y^*_{jt-1} = c_{t-1} - (1-\gamma)c_{t-2} + e_{jt-1} - e_{j,t-2} - \gamma c_{t-3}$$

We now see that Δy^*_{it} on Δy^*_{jt-1} share two common terms: c_{t-1}, which appears with opposite signs ($-(1-\gamma)$ and $+1$); and c_{t-2} which appears with the same sign ($-\gamma$ and $-(1-\gamma)$). It follows that the negative correlation between Δy^*_{it} on Δy^*_{jt-1} *falls* as γ increases, and disappears if $\gamma = 1$. In other words, when common shocks affect the benefits from usage over two periods, the negative correlation between Δy^*_{it} on Δy^*_{jt-1} tends to disappear. Simulations show that things are slightly different for Δy_{it} on Δy_{jt-1}: the negative correlation between them also falls as γ tends to 1; but it never fully disappears and remains negative throughout. We suspect that this arises due to the fact that the transformation $1[y^*_{it} > \tau]$ is non-linear, which means that the covariance between Δy_{it} on Δy_{jt-1} is not additively separable in c_t and e_{it}. The main conclusion remains, however: it is not possible to explain a positive correlation between Δy_{it} on Δy_{jt-1} by using contemporaneous or two-period common shocks. A negative correlation between them is, however, compatible with the existence of such common shocks.

With two-period common shocks, there is a negative correlation between Δy^*_{it} and Δy^*_{jt-2} (see equation B7) which increases with γ. But, as is clear from equation (B8) there is no correlation between Δy^*_{it} and Δy^*_{jt-3}.

$$\Delta y^*_{jt-2} = c_{t-2} - (1-\gamma)c_{t-3} + e_{jt-2} - e_{j,t-3} - \gamma c_{t-4} \tag{B7}$$

$$\Delta y^*_{jt-3} = c_{t-3} - (1-\gamma)c_{t-4} + e_{jt-3} - e_{j,t-4} - \gamma c_{t-5} \tag{B8}$$

Common shocks and adoption

In the specific regression format used in our paper, the regressor of interest is the cumulative adoption by neighbors, which is denoted as A_{it}, and the main regressions of interest regress Δy_{it} on the lagged first-difference in A_{it}. Two main regressions are estimated, depending on whether $z_{it} = 0$ (not adopted yet) or $z_{it} = 1$ (used at least once). We also allow for growth in the size of i's neighborhood of potential adopters. All these modifications are introduced because they correspond to the predictions regarding adoption and usage derived from the model presented in the paper, and they are essential to our testing strategy. Here we ask what these modifications imply for estimated coefficients of interest in the presence of common shocks to y^*_{it}.

We start by noting that the observations made so far extend to the regressions estimated on usage, that is, for $z_{it} = 1$. Unlike Δy_{jt-1}, ΔA_{jt-1} is never negative. It remains, however, that common shock c_{t-1} affects adoption by neighbors and therefore enters the construction of both Δy_{it} and ΔA_{jt-1}, albeit with a different sign. This generates the same prediction of a negative correlation between Δy_{it} and ΔA_{jt-1} in the presence of common shocks.

The situation is different in the adoption regression of Δy_{it} on ΔA_{jt-1} where we only use observations for which $z_{it-1} = 0$, meaning that ME2U had not yet been adopted by i in the

previous period. In these adoption regressions, we find that, when we allow for two-period common shocks (i.e., $\gamma > 0$), we obtain, on average, a positive coefficient on ΔA_{jt-1}. This difference is due to the fact that we only consider positive changes, i.e., situations in which $\Delta y_{it} = \{0, 1\}$; negative changes are omitted. This has the mechanical effect of dropping all observations involving $\Delta y_{it} = -1$, thereby curtailing any negative correlation with ΔA_{jt-1} which, by construction, only takes values $\{0, 1\}$ as well.

The important thing to note, however, is that, in both the adoption and usage regressions, Δy_{it+1} on ΔA_{jt-1} are uncorrelated in the presence of contemporaneous common shocks, and that Δy_{it+1} on ΔA_{jt-2} are uncorrelated in the presence of two-period common shocks.

Appendix C. Assessing the bias arising from unobservable selection

Oster (2019) shows how the size of the bias posed by unobservable selection, under certain assumptions, can be inferred from coefficient and R-squared differences across models with different sets of control variables. Adopting Oster's notation, let the parameter δ denote the proportional selection relationship. If unobservable and observable factors are equally related to treatment, $\delta = 1$; if unobservable are more strongly related to treatment than observable factors, $\delta > 1$; and if observable factors are more strongly related to treatment than observables, $\delta < 1$. Further, let $R_{\max}$ denote the R-squared from a hypothetical regression of the dependent variable on the treatment variable and the observable and unobservable determinants of the dependent variable. For a model that is linear in a single treatment variable, Oster shows how the bias on the treatment coefficient obtained from a regression where observable but not unobservable factors are included can be written as approximately equal to $\delta \left[\beta^0 - \widetilde{\beta}\right] \frac{\left[R_{\max} - \widetilde{R}\right]}{\widetilde{R} - R^0}$, where β^0 denotes the coefficient resulting from the short regression of the dependent variable on the treatment variable with observable control variables excluded; R^0 is the R-squared from the short regression; $\widetilde{\beta}$ is the coefficient from the regression with observable control variables included, and $\widetilde{R}$ is the R-squared from that regression. Clearly, the bias in $\widetilde{\beta}$ can be severe if: unobservable factors are strongly related to treatment (in which case δ is high); if the treatment coefficient changes considerably as a result of the addition of observable control variables (in which case $\left[\beta^0 - \widetilde{\beta}\right]$ is high) while at the same time the R-squared doesn't move much (in which case $\widetilde{R} - R^0$ is low); and/or if the unobservable factors (would) have considerable explanatory power (in which case $R_{\max} - \widetilde{R}$ is high). Of course, neither δ nor $R_{\max}$ is observable, but the bias formula above is nevertheless useful as it enables researchers to quantify the bias for specific values of δ and $R_{\max}$. Clearly, if there is no movement in the treatment coefficient as we move from the short regression to the regression with observable controls included, Oster's framework

implies that there is no bias, regardless of the values of δ and $R_{\max}$.

Results are shown in Table C1.

References

Oster, Emily (2019). "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 37(2), 187-204.

Table C.1

First Adoption: Robustness to selection on unobservables

	(1) Linear model: Partially controlled	(2) Linear model: Fully controlled	(3) Bias adjusted β	(4) Bias adjusted β	(5) δ for $\beta=0$
ΔA(i,t-2)	0.0041	0.0039	0.0031	0.0023	0
s.e.	0.0005	0.0005	--	--	--
R-squared	0.031	0.038			
δ			1.0	2.0	4.1
R_{max}			0.077	0.077	0.077
Controls					
ΔS(it)^2	Y	Y			
Year x month	Y	N			
District	Y	N			
Cell tower	Y	Y			
Year x month x district	N	Y			
Observations	87,563	87,563			

Note: Columns (1) and (2) show results for a linear specification of the form Δy(i,t+1) = β*ΔA(i,t-2) + controls + Δε(i,t+1). Standard errors are clustered at the district level (M=27). Columns (3)-(4) show bias-adjusted estimates of β, based on the approach developed by Oster (2019). Column (5) shows the value of δ for which β = 0, again based on Oster (2019). Oster's approach is not suitable for specifications where the potentially endogenous explanatory variable enters nonlinearly (as in Table 2), hence we consider linear specifications for the analysis of robustness to selection on unobservables.

Appendix D. Robustness analysis

In this part of the appendix we show regression results for specifications corresponding to those in Tables 2, 3 and 4 in the main paper, with alternative lag lengths for the explanatory variable ΔA.

Table D1. First Adoption: First Difference Estimates

	(1)		(2)		(3)	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
ΔA(i,t-1)	0.000736	0.00067	0.000735	0.000701	0.00062	0.000654
ΔS(it)^2	6.09E-05	3.63E-05	-0.00075	6.04E-05	-0.00073	6.61E-05
ΔA(i,t-1)^2	-3.19E-05	3.12E-06	-3.1E-05	3.85E-06	-3E-05	3.54E-06
Δ[A(i,t-1)S(it)]	0.000291	1.75E-05	0.000308	3.01E-05	0.000301	3.07E-05
R-squared	0.006004		0.035255		0.042274	
Observations	91826		91826		91826	
Marginal effect of A(i,t-1) at means of A(i,t-1) and S(it)						
A(i,t) = sample mean	0.004788	0.000576	0.00515	0.000562	0.004961	0.00052
Marginal effects of A(i,t-1), at different levels of A(i,t-1)						
A(i,t-1) = 0	0.006307	0.000697	0.006644	0.000695	0.006392	0.000649
A(i,t-1) = 20	0.005034	0.000594	0.005392	0.000582	0.005192	0.00054
A(i,t-1) = 40	0.00376	0.000502	0.00414	0.000491	0.003993	0.000448
A(i,t-1) = 60	0.002486	0.000426	0.002888	0.000436	0.002794	0.000388
A(i,t-1) = 80	0.001212	0.000378	0.001636	0.000433	0.001594	0.000374

Note: The dependent variable is Δy(i,t+1). Standard errors are clustered at the district level (M=27). Marginal effects are evaluated at sample means of regressors (in levels). *** p<0.01 ** p<0.05 * p<0.10

Table D2. Generalized First Adoption Model: First Difference Estimates

	(1)		(2)		(3)	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
ΔA(i,t-1)	0.003657	0.000788	0.002849	0.000716	0.0027	0.000652
ΔS(it)^2	0.000464	4.66E-05	-0.00024	5.71E-05	-0.00028	5.68E-05
ΔA(i,t-1)^2	-8.45E-06	1.06E-05	-8.00E-06	9.76E-06	-7.65E-06	9.63E-06
Δ[A(i,t-1)S(it)]	2.04E-05	3.65E-05	5.9E-05	4.05E-05	6.29E-05	4.18E-05
Δ[m(it) x S(it)]	0.061519	0.004524				
Δ[m(it) x A(i,t-1)]	-0.007956	0.00161	-0.00661	0.001708	-0.00619	0.001789
Δ[m(it) x S(it)^2]	-0.000968	5.79E-05	-0.00126	0.000159	-0.00112	0.000203
Δ[m(it) x A(i,t-1)^2]	-2.78E-05	1.19E-05	-3.6E-05	1.4E-05	-3.6E-05	1.37E-05
Δ[m(it) x A(i,t-1) x S(it)]	0.000405	5.08E-05	0.000476	7.51E-05	0.000461	7.81E-05
R-squared	0.010074		0.047223		0.059898	
Observations	90584		90584		90584	
Marginal effect of A(i,t-1) at means of A(i,t-1) and S(it)						
m(it) = 0	0.00366	0.000843	0.003509	0.000787	0.003438	0.000748
m(it) = 1	0.005411	0.001169	0.008555	0.001386	0.008521	0.001377
Marginal effects difference[(a)]	-0.001751	0.001393	-0.00505	0.001571	-0.00508	0.001555

Note: The dependent variable is Δy(i,t+1). Standard errors are clustered at the district level (M=27). Marginal effects are evaluated at sample means of regressors (in levels). Datapoints for which Δm(it)=1 (i.e. where m(it) switches from 0 to 1) are excluded for these estimations. Δ[m(it) x S(it)] is collinear with the fixed effects in (2) and (3), and is therefore excluded from these specifications. *** p<0.01 ** p<0.05 * p<0.10

(a) This is equal to the marginal effect at m(it)=0 minus the marginal effect at m(it)=1.

Table D3. Adoption & subsequent usage: First Difference Estimates

	(1)		(2)		(3)	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
ΔA(i,t-1)	0.000736	0.00067	0.000735	0.000701	0.00062	0.000653
ΔS(it)^2	6.09E-05	3.63E-05	-0.00075	6.03E-05	-0.00073	6.6E-05
ΔA(i,t-1)^2	-3.19E-05	3.12E-06	-3.1E-05	3.85E-06	-3E-05	3.53E-06
Δ[S(it) x A(i,t-1)]	0.000291	1.75E-05	0.000308	3.01E-05	0.000301	3.06E-05
Δ[z(it) x S(it)]	-0.033173	0.001357				
Δ[z(it) x A(i,t-1)]	-0.003498	0.000843	-0.00336	0.000832	-0.00328	0.000791
Δ[z(it) x S(it)^2]	-1.49E-05	3.34E-05	0.000806	6.24E-05	0.00079	6.8E-05
Δ[z(it) x A(i,t-1)^2]	3.44E-05	3.55E-06	3.37E-05	4.20E-06	3.23E-05	3.90E-06
Δ[z(it) x A(i,t-1) x S(it)]	-0.000294	1.86E-05	-0.00031	3.01E-05	-0.0003	3.04E-05
R-squared	0.00538		0.009474		0.011005	
Observations	362485		362485		362485	
Marginal effect of A(i,t-1) at means of A(i,t-1) and S(it)						
z(it) = 0	0.004788	0.000576	0.00515	0.000562	0.004961	0.000519
z(it) = 1	-0.00247	0.000352	-0.00224	0.000368	-0.00226	0.000364
Marginal effects difference[(a)]	0.007258	0.000741	0.007388	0.00071	0.007216	0.000677

Note: The dependent variable is Δy(i,t+1). Standard errors are clustered at the district level (M=27). Marginal effects are evaluated at sample means of regressors (in levels). Datapoints for which Δz(it)=1 and Δz(i,t-1)=1 (i.e. the period when z(it) switches from 0 to 1 and the subsequent period) are excluded for these estimations. Δ[z(it) x S(it)] is collinear with the fixed effects in (2) and (3), and is therefore excluded from these specifications. *** p<0.01 ** p<0.05 * p<0.10.

(a) This is equal to the marginal effect at z(it)=0 minus the marginal effect at z(it)=1.

Table D4. First Adoption: First Difference Estimates

	(1)		(2)		(3)	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
ΔA(i,t-3)	-0.00321	0.000535	-0.00152	0.000639	-0.00163	0.000623
ΔS(it)^2	2.52E-06	4.32E-05	-0.0007	6.79E-05	-0.00067	7.48E-05
ΔA(i,t-3)^2	-2.8E-05	3.56E-06	-2.8E-05	4.89E-06	-2.7E-05	4.77E-06
Δ[A(i,t-3)S(it)]	0.000325	2.51E-05	0.000312	4.08E-05	0.000304	4.26E-05
R-squared	0.00451		0.031497		0.038791	
Observations	83480		83480		83480	
Marginal effect of A(i,t-3) at means of A(i,t-3) and S(it)						
A(i,t) = sample mean	0.001708	0.000418	0.003159	0.00041	0.002962	0.000417
Marginal effects of A(i,t-3), at different levels of A(i,t-3)						
A(i,t-3) = 0	0.003009	0.000499	0.004458	0.00056	0.004194	0.000574
A(i,t-3) = 20	0.001878	0.000426	0.003328	0.000426	0.003123	0.000435
A(i,t-3) = 40	0.000747	0.000392	0.002199	0.000354	0.002052	0.000348
A(i,t-3) = 60	-0.00038	0.000409	0.001069	0.000381	0.000981	0.000355
A(i,t-3) = 80	-0.00152	0.00047	-6.1E-05	0.000492	-9E-05	0.000451

Note: The dependent variable is Δy(i,t+1). Standard errors are clustered at the district level (M=27). Marginal effects are evaluated at sample means of regressors (in levels). *** p<0.01 ** p<0.05 * p<0.10

Table D5. Generalized First Adoption Model: First Difference Estimates

	(1)		(2)		(3)	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
ΔA(i,t-3)	6.78E-05	0.000531	0.001217	0.000554	0.001013	0.00056
ΔS(it)^2	0.00042	3.51E-05	-0.00015	6.31E-05	-0.00019	6.62E-05
ΔA(i,t-3)^2	-3.43E-08	6.74E-06	-1.02E-06	7.20E-06	-5.41E-07	7.23E-06
Δ[A(i,t-3)S(it)]	4.67E-05	2.71E-05	5.01E-05	3.34E-05	5.42E-05	3.64E-05
Δ[m(it) x S(it)]	0.067199	0.005855				
Δ[m(it) x A(i,t-3)]	-0.00823	0.001933	-0.00838	0.002002	-0.00821	0.001955
Δ[m(it) x S(it)^2]	-0.00099	4.93E-05	-0.00135	0.000199	-0.00122	0.000244
Δ[m(it) x A(i,t-3)^2]	-3.2E-05	1.07E-05	-3.9E-05	1.41E-05	-3.9E-05	1.41E-05
Δ[m(it) x A(i,t-3) x S(it)]	0.000388	4.87E-05	0.000486	8.52E-05	0.000479	8.9E-05
R-squared	0.008636		0.056096		0.068506	
Observations	82353		82353		82353	
Marginal effect of A(i,t-3) at means of A(i,t-3) and S(it)						
m(it) = 0	0.000838	0.000435	0.002006	0.000436	0.001888	0.000427
m(it) = 1	0.002437	0.001456	0.005852	0.001457	0.005747	0.001487
Marginal effects difference[(a)]	-0.0016	0.001582	-0.00385	0.001633	-0.00386	0.001643

Note: The dependent variable is Δy(i,t+1). Standard errors are clustered at the district level (M=27). Marginal effects are evaluated at sample means of regressors (in levels). Datapoints for which Δm(it)=1 (i.e. where m(it) switches from 0 to 1) are excluded for these estimations. Δ[m(it) x S(it)] is collinear with the fixed effects in (2) and (3), and is therefore excluded from these specifications. *** p<0.01 ** p<0.05 * p<0.10

(a) This is equal to the marginal effect at m(it)=0 minus the marginal effect at m(it)=1.

Table D6. Adoption & subsequent usage: First Difference Estimates

	(1)		(2)		(3)	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
ΔA(i,t-3)	-0.00321	0.000535	-0.00152	0.000639	-0.00163	0.000622
ΔS(it)^2	2.52E-06	4.32E-05	-0.0007	6.79E-05	-0.00067	7.46E-05
ΔA(i,t-3)^2	-2.8E-05	3.56E-06	-2.8E-05	4.88E-06	-2.7E-05	4.76E-06
Δ[S(it) x A(i,t-3)]	0.000325	2.51E-05	0.000312	4.07E-05	0.000304	4.26E-05
Δ[z(it) x S(it)]	-0.04198	0.001965				
Δ[z(it) x A(i,t-3)]	0.001589	0.00091	0.000224	0.001079	0.000347	0.001036
Δ[z(it) x S(it)^2]	4.93E-05	4.09E-05	0.000755	6.64E-05	0.000731	7.37E-05
Δ[z(it) x A(i,t-3)^2]	3.01E-05	3.83E-06	2.98E-05	5.19E-06	2.83E-05	5.07E-06
Δ[z(it) x A(i,t-3) x S(it)]	-0.00033	0.000026	-0.00032	4.24E-05	-0.00031	4.41E-05
R-squared	0.005171		0.008884		0.010396	
Observations	353406		353406		353406	
Marginal effect of A(i,t-3) at means of A(i,t-3) and S(it)						
z(it) = 0	0.001708	0.000418	0.003159	0.000409	0.002962	0.000416
z(it) = 1	-0.00168	0.00048	-0.00128	0.000473	-0.00126	0.000473
Marginal effects difference[(a)]	0.003389	0.000577	0.004441	0.000574	0.004225	0.000528

Note: The dependent variable is Δy(i,t+1). Standard errors are clustered at the district level (M=27). Marginal effects are evaluated at sample means of regressors (in levels). Datapoints for which Δz(it)=1 and Δz(i,t-1)=1 (i.e. the period when z(it) switches from 0 to 1 and the subsequent period) are excluded for these estimations. Δ[z(it) x S(it)] is collinear with the fixed effects in (2) and (3), and is therefore excluded from these specifications. *** p<0.01 ** p<0.05 * p<0.10.

(a) This is equal to the marginal effect at z(it)=0 minus the marginal effect at z(it)=1.