[SQUEAKING]

[RUSTLING]

[CLICKING]

GLENN ELLISON: OK, let me go ahead and get started. Today's topic is going to be entry. And I think entry is an incredibly important topic in IO. In fact, one of the messages I'm going to talk about today is, in some sense, just about everything we've done in this semester so far is kind of like rounding error compared to what we do now.

So how do I think about entry? So far, we've talked about a lot of models that have the number of firms fixed. And I've talked about lots of things about factors that affect pricing with the number of firms fixed. You know, monopoly pricing, if you have one firm, how does that-- what do they do? And then static competition with vertical differentiation or horizontal differentiation or search costs or other things, and it's always the nature of the market is determining what's the level of markups and what inefficiencies there might be.

But that's all taking market structure as fixed. And so starting today, I'm going to think about what happens if you allow the number of firms to change. And so the simplest way to think about this is let's look at a two-stage model. I'm going to start with the case of homogeneous goods. So you can think of this as old-fashioned steel plants entering in each-- producing cold-rolled steel or aluminum plants or some industry that has a small number of firms because they're big fixed costs of setting up plants and then they compete with each other.

So stage one, we have a large number of potential entrants. The potential entrants choose in or out. If they decide to come in, they pay some fixed entry cost K. And I'm going to think about this, most of the lecture, as a sunk cost. That is, once you build a steel plant, it's a steel plant, you can't use it for anything else. You're not going be able to sell it to somebody else, do something else with it.

Firms that choose to enter then play some game like Bertrand competition, Cournot competition, price competition with search, et cetera. And assume firms have some variable costs c of Q. And then, how are you going to solve this model? I'm going to solve this model by backward induction. I first start with the second stage. In the second stage, if there are N firms, there are going to be some equilibrium prices, p star of N. p star of N times Q star of N minus c of Q star of N is going to be the variable profits of the firms if they enter.

And then you find the number of equilibrium entrants by looking at the first stage. And what's that going to be? OK, first comment, I guess these slides were made up in the COVID era, where I would write on them right on my iPad. But let me go ahead and do this on the board. So let me graph what our profits look like as a function of the number of firms.

The easiest one to graph is Bertrand. So 1, 2, 3, 4, 5. Bertrand competition, you have this level of profits, pi M, if you have one firm. And then as soon as you have two firms, profits drop to 0.

Another thing you could have is Cournot competition. If you have Cournot competition, you get exactly the same outcome when N equals 1. But then profits drop when you add 2, 3, 4, or 5, whatever firms. If you think about-- if we did Cournot with linear demand-- Cournot with linear demands, you were getting, like, P star is approximately c plus 1 over--

Or simplest case-- simplest case for Cournot was N firms demand P as 1 over P. And then I think in that model, it was P star was equal to 1 over N plus 1. And q star was equal to 1 over N plus 1. So profits were like 1 over N squared. So Cournot profits, approximately 1 over N squared or 1 over N plus 1 squared, whatever.

Another model I did was a model with search. In a model with search, if you remember, there was the Stahl model, you would always have the upper bound of the distribution, always gave you the monopoly profits as long as the search costs were in an appropriate range. And so that one declined something-- if you think about that, that means it declines slower because the firms are serving the nonshoppers.

So in the search model, you're serving the nonshoppers. You're always indifferent to charging the monopoly price. And so you get the monopoly price divided by N plus-- divided by N times 1 minus mu is the profits. And they go down like 1 over N. But whatever you have, what's going to happen in this model in the first stage-- so if you have some level of fixed-- let me focus on Cournot.

Let's suppose you do Cournot, and you have some fixed level of entry cost K. Then, in equilibrium, you're going to have exactly two firms enter. Because when two firms enter, the firms' profits are above K. If you had three firms in the market, the profits would be below K, so N equals 2. And so the number of firms, it's locally invariant to K because you can raise or lower costs, and it doesn't affect things. Obviously, if K drops, then the third firm enters.

(That's a terrible piece of chalk.) You could also think of a mixed strategy equilibrium solution. If there's some large but finite number-- if there's some number of M of potential entrants, where M is much bigger than N star-- you could also think of M firms each mixing in or out, and then you get a mixed strategy equilibrium.

In the mixed strategy equilibrium, all the firms that are mixing are going to have to have expected profits exactly equal to K. And so in that model, all the firms that are mixing have expected profits exactly equal to K. Therefore, here you can think of it as you're going to expect to get two-plus firms, between two and three firms in the market. But if it's-- obviously, if it's like Poisson draws, sometimes you'll get one, two, three, four, five, or more.

In models like a procurement auction, where there's one date on which everyone has to submit sealed bids and no one knows how many other bidders there are, this expected pi of N equals K is probably the right approach. If you have models like the steel plants where the firms are not entering simultaneously, they're entering sequentially, then it's probably like the pure strategy equilibrium is the one you want to focus on because no one's going to open a third steel plant if the third steel plant is going to drive prices below where you need to cover the fixed costs.

So observations-- fixed costs and competition determine the number of entries, number of entrants. But with lower fixed costs, we get more entry. In many models, the number of firms goes to infinity as the fixed costs of entry go to 0. The only time that's not going to happen is if you had-- I talked about the low demand models, where price asymptotes to a positive-- no, it's still going to work. So where are you not going to N star--

You're not going to get N star goes to infinity-- like, in Cournot, you don't get N star goes to infinity because in Cournot, in fact, you get N star is 1. So if profits hit 0 at some point before, when N gets large, and even with no entry costs, you can still get-- or small positive entry costs, you can still get finite entry.

Profits are, in this model, are a rounding error. And we spent all this time working on markups, but here, the profits are just-- if this happens to be the fixed cost, you get pretty big profits. If this was the fixed cost, you'd get very little profits. If the fixed costs dropped here, you'd get even higher profits. When then they drop more, you get hardly any profits at all.

So equilibrium profits that we observed really don't have to be related to anything we've talked about so far because all that matters is how many firms fit in the market. And then everything else just determines where you are in between the profits with two firms or three firms or three firms and four firms. And so profits just become this rounding error that's just coincidence of where entry costs big enough that the last firm just barely gets in.

Now, that's some-- that is equilibrium profits taking the fixed costs of entry into account. It's still the case that the markups are higher when you have smaller numbers of firms. The markups are higher, but it's just the markups just offset the fixed costs the firms are incurring.

In many models, the equilibrium firm size increases when the market size increases. So I give the Cournot example here. In the Cournot example, suppose you have little m consumers in the market, and so demand becomes m times D of p. Let's suppose we double the market size. If you double the market size and you double the number of firms, then the number of customers per firm would be the same. But there's a second effect of the prices decreasing.

And so if the prices decrease-- the prices decrease, then the firms are worse off. So Cournot, in some sense, it's like the profits are going down like 1 over N squared. With N firms, profits are going up proportional to m if the number of customers goes up. So you're going to get-- firm growth is going to be the square root of the number of customers. So as the firm-- as the market increases, firms go up square root of the number of customers goes up. So firm size goes up like square root of the number of customers.

Any questions on these simple observations? First substantial topic I want to talk about today is, what's the efficiency of entry? Do we get more firms and markets than we want? Do we get fewer firms and markets than we want? And this is an important practical question. There are many, many industries and many, many countries where the firms have gotten together and have some kind of government-sanctioned entry limitations.

There's a question on the Massachusetts ballot about should liquor stores-- or should firms be able to sell beer in more than nine stores in the state of Massachusetts? For many years, you were only allowed to sell beer, I think, in one store in the state of Massachusetts. So all the supermarkets, you notice-- I don't know if-- if you go to-- supermarkets in other states have beer. Supermarkets in Massachusetts typically didn't have beer because there was a law that said you could only sell beer in one of your stores in the entire state of Massachusetts.

And so Star Market had to pick one Star Market to have beer. And now there are-- now you can sell beer in nine Star Markets, and they're thinking of increasing that level. But why do we get these entry laws? Well, the firms that are in the market always argue that competition entry is wasteful, and we need to save the mom and pop stores and the small businesses and protect them from, quote, excessive entry.

So anyway, here's a discussion of do we think entry should or shouldn't, would or would not be excessive in markets? What's the first best? Here I'm going to do the homogeneous case first. Homogeneous goods, there's no idiosyncratic preferences. So the first best is just one firm-- have that one firm price at marginal cost. That's typically not feasible because you'd need price regulation to force the monopolist to price at marginal cost.

So what we often compare it to is the second best, where the second best you take p star of N as given. You can't affect the nature of the competition, and you just say, what would be the optimal number of firms, given that p star of N can be determined in equilibrium? And so this would be our welfare calculation. The welfare with N firms is equal to the integral 0 to N-- did we do big Q or little q? Little q.

So if you have N firms, this is the gross surplus produced by firms producing total quantity N q star of N. This is the total production costs incurred when each of N firms produces quantity q star of N. And these are the fixed costs. So we want to maximize that. So the first theorem is that in this model with homogeneous goods, entry is always excessive. So we do get too many firms and markets. We would like to restrict it.

Why? Here's a theorem. So suppose that this welfare function is concave. I know I've talked about how things are often not concave. Welfare functions here would be concave because as you get lots of firms, you just asymptote to line of slope K. It's inefficient to keep paying these fixed entry costs over and over and over again. So suppose the welfare function is concave.

Suppose the variable profits are decreasing in number of firms N. And suppose that the equilibrium price per firm quantities can be extended to differentiable functions that have these natural properties. As the number of firms increases, total output goes up. As the number of firms goes up, each individual firm's output goes down. And at the equilibrium number of firms, price minus marginal cost is positive, which it always should be.

Sorry, this is given any number of firms-- given any number of firms N, the resulting price would be above marginal cost. The theorem is that then N is at least N second best minus 1. So it's entries excessive, although it can, due to rounding errors, be one too few. Bertrand is a case where it's one too few. Sometimes you'd like two firms, you only get one. But the general result is that you get too many firms, other than this possible rounding error issue.

So proof. So first, we've got this welfare function. The welfare function is naturally described only at a set of integer points. But in many of these models, like Cournot over there, you have p star of N equals 1 over N plus 1, Q star of N equals 1 over N plus 1.

So you can just naturally treat this as a continuous function when N is a continuous variable, like use a function Q star of x equals 1 over x plus 1. And by doing that, you extend the-- you just draw a smooth curve through those welfare points.

I'm going to let N hat second best be the solution to where this thing's first order condition is 0. So that would be right here. N hat second best. So I define N hat second best to be the maximizer of this smooth curve. The second best is always going to be either-- if W is a concave function, it's either going to be this point or this point. So the optimal number of firms is always at most what you get-- optimal number first is always at most what you get by just rounding up this point, N hat second best.

So what is the first order condition defining N hat second best? Well, it's just-- the first order condition is d W of N dN equals evaluated at N hat second best equals 0. That's what determines this point here.

What does that derivative look like? So I differentiate this expression with respect to N. The first term only depends on N through the upper limit of integration. So the derivative of this with respect to N is just the value of this evaluated at N q star of N, P of N q star of N times the amount by which the upper limit moves when N changes. So the derivative of this with respect to N is what I wrote down here.

The second effect of N on welfare is this term changes, Nc of q star of N. So the derivative of that with respect to N is c of q star of N plus Nc prime of q star of N, q star prime at N. So that's a derivative of the second term. And then the derivative of the third term with respect to N is just K.

So I just differentiate this expression with respect to N. I say that evaluated at N hat second best is going to be equal to 0. And then I'm going to start grouping terms. The first term I pull together are P times q minus c of q. And so I'm going to take this term times this minus this and call that the variable profit. So that's pi v star of N.

What's left over? Everything that's left over has an Nq star prime of N. Well, there are two terms left over. This times this and this that have an Nq star prime at N. And those become Nq star prime at N times price minus marginal cost. And then the third thing that's left over is the K at the end.

And then under this assumption that I made that price minus marginal cost is always positive at the equilibrium prices, this is a positive number. And so what that means is that if you evaluate this at N hat second best, I get variable profit minus K plus this equals 0.

That tells me that the variable profit evaluated at N second best minus K is strictly positive, which means that 1x-- at N hat second best, variable profit is bigger than K. So that means other firms would like to come in. So if we had exactly this many firms, not that this is an integer, the next firm always wants to come in. So that means the equilibrium entry-- oops. So equilibrium entry is at least here.

Because at this point, we know a firm wants to come in. At this point, we don't know. But if you're less, then you definitely want to come in because pi v star is a decreasing function. So pi v star is bigger than K at this point. So pi v star is also bigger than K at this point. So N star is somewhere out here. So N star is at least what you get-- is at least here.

And what did I say? I said that the first best and second best is one of these. So if N second best is either this point or this point, the equilibrium number is somewhere this point or higher, we're always getting too many firms in. Is that OK?

I have on the next page easier, maybe simpler intuition. In a homogeneous good model, the only benefit that anyone gets from entry is that we reduce the deadweight loss. We get quantity increases. So what are the total gross surplus, total consumer and firm benefits from the extra entry? It's just the extra quantity that gets produced times p star minus c, where p star is like the equilibrium profit you get at the equilibrium number of firms.

So when we let in more firms, the welfare benefit is just, how much did quantity go up times p star minus c? What does the entering firm get? The entering firm gets much more than this because it's not-- the entering firm is not just selling the amount by which quantity has gone up. The entering firm is getting that quantity-- it's getting the extra quantity, but it's also stealing some quantity away from all the other firms that were there before.

And so people call that the business stealing effect, that the amount by which quantity goes up, which is the welfare gain, is much smaller than the amount by which sales go up for the last firm that comes in. So the last firm's profits-- the social cost of it coming in is K. The social benefit of it coming in is like delta q times p minus c. But its profits are much bigger than-- its variable profits are much bigger than delta q times p minus c because it's getting all the profits and all the other units it's taking away from other firms instead. And so that's the business stealing effect.

And in this homogeneous good model, that business stealing effect is positive. Therefore, you get too much entry other than this possible rounding error question.

As I said, it's true up to a rounding error, but if you do numerical examples, you find that you often get many, many more firms than you'd want socially. And the reason is that it's not just some of the demand is business stealing. Most of the demand is business stealing. So in this Cournot example, D of p is 1 minus p. q star of N is 1 over N plus 1.

So what do you get from adding an Nth firm? When you have N firms production goes up to N over plus 1 where it used to be N minus 1 over N, so that difference is 1 over N times N plus 1. The Nth firm is producing 1 over N plus 1. So 1 over N fraction of its output is actual added output. N minus 1 over N of its output is business stealing output.

And so in some sense, the incentive that it has to enter is N times too large because it's treating 1 over N plus 1 as the benefit it's getting, whereas 1 over N times that is actually the social benefit that it's producing. And so because private benefit is N times the social benefit-- when private benefit is N times the social benefit, you can get way, way too many firms into the market.

OK, so now let me think of it-- that was entry when you have homogeneous goods. What happens when you get away from homogeneous goods? When you get away from homogeneous goods, entry can be too high or too low. And why is that? Well, the general setup-- welfare when you have N firms, it's going to be the sum of all the firms in the market of their variable profits plus the total consumer surplus in the market minus the fixed costs.

If I add an N plus first firm, I get the sum over N plus 1 firms instead of N firms of the variable profits. I now get consumer surplus when you have N plus 1 firms, but I pay higher fixed costs. So what's the change in welfare from the last entrant? First, let me group together the firm that's here that wasn't in this sum, the firm N plus 1. I get the variable profit of firm N plus 1 minus its fixed cost. That's a difference here.

And then I get two other terms. The first is subtracting the corresponding i's here, I get sum i equals 1 to N variable profits of firm i with N plus 1 versus N firms. And then I get the consumer surplus with N plus 1 minus N firms.

So when I wrote here "entry is socially optimal," I probably should have said entry of firm N plus 1 is socially optimal if this delta W is positive. Why might free entry-- so what we want is if this is positive, we want firm N plus 1 to come in. And if firm-- if delta W was negative, we don't want firm N plus 1 to come in.

So first of all, why might entry be excessive? So entry would be excessive if this was positive but this was negative. Why could this be positive and this be negative? Well, if this is positive but small to make a firm want to come in but this is a big negative number-- and why could this be a big negative number? This is how every other firm's profits are affected by me coming in.

If I drive down prices and I take quantity away from the other firms, they suffer both because the profits they earn on the units they still sell go down and because they don't sell some units and they don't earn p star minus c. So this is what we call-- again, this is the business stealing effect. If this business stealing number is always negative-- if it's a big negative number and this is a small positive number, we can get the entrant coming in, even though this business stealing term is negative.

But then the second effect is that we could have firm N doesn't come in because this is negative. And this is negative but delta W is positive because there's a big increase in consumer surplus. And I showed the homogeneous good model this doesn't happen except with that rounding error effect.

But once you get to heterogeneous goods, it could be that when you get more goods in the market, consumer surplus goes up by more than prices go down because consumers get a good that's better matched to their tastes. And because the new good is better matched to their tastes, there's a big increase in consumer surplus that isn't just the loss and profits.

And so if the increase in consumer surplus is coming from other channels like match quality rather than just from prices going down, then this term can outweigh the business stealing effect and you could have too few firms in a market because firms-- firms don't like-- I mean, firms not internalizing consumer surplus is why you get too few firms in the Bertrand model. In the Bertrand model, firms don't realize I come in, yes, I lose some money, but all those consumers are way better off because the price drops so much.

But in another model, it's the consumers don't realize-- the firms don't realize that, like, I come in, I don't make that much money, but consumers are so happy to have my product that social welfare is better. OK, any questions?

So in a general model, the most general conclusion is that entry can be too high or too low socially. And it depends on which of these effects is more important, the social welfare effect or the business stealing effect. Prices going down alone is not enough to do it. You've got to have other channels of the consumer surplus going up.

I just figured I'd put in one slide of what happens if I actually compute these things. I would say that if you write down the most obvious models you're going to write down to try to compute these things. While this consumer surplus effect means that welfare can-- we can get too few firms, most models you write down, if you're not thinking about writing down something to get it to happen, you're going to also conclude that there's still way too many firms in markets.

So I do an example here. So consider a Hotelling-like model, where you have a mass m of consumers. And I'll write over here. So let's think about a Hotelling circle. So this is, ah-- ice cream stands locating around like a Hawaiian island that has just beaches around the entire perimeter. And so if you get N firms-- suppose you get N firms; they locate themselves evenly around the circle.

In the c plus t model-- if you have N firms, the transportation cost distance between here and here is only going to be t over N when you have N firms. So if it's t to walk the whole perimeter of the island, it's t over N when you put N firms around the perimeter. So if you get N firms that are arranged evenly, what you basically get is each firm here is just competing with the firm on its right and the firm on its left.

I didn't draw those so evenly, but let's assume this distance is t over N, this distance is t over N. So it's just like I'm playing two simultaneous Hotelling games, one against this firm, one against this firm. So the equilibrium becomes p star equals c plus t over N.

What's the variable profit? If there are m consumers on the beach, I get a fraction 1 over N of them in equilibrium, and my markup is t over N. So my profits are mt over N squared. I set that equal to K, and then it just becomes N equals square root of m times t over K.

So again, as I said, this is, again, a larger firms theorem that says as the market size goes up, the number of firms goes up in this model, also like the square root of the number of customers.

What's the social welfare? There are m consumers. So suppose we're in this model and these things are both t over N. Social welfare is just total gross surplus received minus costs of production minus average mismatch disutility or transportation cost disutility.

This guy here sells to customers who are within distance t over 2N to the left and customers within t over 2N to the right. So he's selling to customers who are-- t over N is this distance. So everybody equidistant between him and the next firm over buys from him. If you sell to customers whose distance ranges from 0 to t over 2N, the average distance that people have to travel is t over 4N.

So this is the extra welfare increase that I mentioned, that as N increases, the distance traveled goes down. That causes a consumer surplus increase that's separate from the one that comes from the prices dropping.

Anyway, so in this model, total social welfare is m times v minus m times c minus m times t over 4N minus NK. Notice that m times v minus c is completely independent of N. And the Hotelling model, assuming v is large, everybody buys all the time.

So what we get is-- differentiating this with respect to N, derivative of this with respect to N is mt over 4 times 1 over N squared. Derivative of this with respect to N is just K. And so what I get is that m over 4 times 1 over N squared minus K, if I add N second best equals 0, N second best is square root of mt over 4K, which is exactly half of N star.

So this is saying, in some sense, in equilibrium, you get exactly twice as many firms as you would like. And this is just-- it's another case that in this model-- so this model has the welfare benefit of people getting better matches. This model has no market expansion effect whatsoever. When you put in an Nth firm, it's 100% business stealing.

And so it turns out that because the market expansion is 100% business stealing, so business stealing is a big deal, this extra benefit that you get from extra match quality isn't nearly large enough to overcome the business stealing negative, and you're getting twice as many firms. And I think, in some sense, I view this as an important cautionary tale. You're going to write down models with logit demand and estimate how many firms you're going to get and how many firms you think are optimal or whatever.

A lot of models like this are set up-- again, I could have done this model, and I could have said I'm estimating demand elasticities by estimating t. But whatever value of t I estimated when I estimated this model, my model-- I do my counterfactual and say how many firms would be socially optimal, and socially optimal would be exactly half the number that are observed. So it's just important to be aware of. Do models have built-in answers as to whether the consumer surplus effect is going to outweigh the business stealing effect or not? And this one, it's totally built in. And the answer is it's a factor of 2.

But I think it's generally true. Most models you write down, you are going to get a factor-- you are going to find that there are way too many firms. And I don't know if I-- maybe I believe that. How would you get-- how would you get bigger social welfare benefits? There's two ways you would get bigger social welfare benefits.

The first is-- in this model, people's utility for getting a product that's not ideal just goes down linearly. If you imagine people's utility looked more like this, so there are people who really, really like having a product that's very close to their-- very close to their ideal point. So when you add an extra and have three ideal points instead of two, you're giving some people lots and lots of extra utility.

So having something like this, people who like things very close to their ideal point, would be one way to get entry to be not too high. Obviously, you can only go so high here before the firm is going to give up on serving everyone and just serve the people who like its product a lot at a really high price. So that's going to be kind of like those information design things where you can only make the curve go up so much and have the firm still sell to everybody.

And then the other effect is here, again, 100% of your demand is business stealing. If you think of an infinite dimensional space, where most people are not buying any product at all, so every new firm that comes in is mostly serving new customers. So if you can make the business stealing effect small and make this consumer surplus large for people who get close matches, those are the things that are going to make entry be insufficient in practice.

And I think if you want to do an empirical study, you'd really want to think about, do I have free parameters to capture what this surplus looks like? And do I have a free parameter to capture how much of the people are being added to the model versus how many people are just business stolen from another firm?

For instance, I read that nested logit has that extra parameter that substitution to the outside good is different from substitution to the inside goods. And a pure logit with no outside good is 100% business stealing model. Questions?

Entry with vertical differentiation. OK, so let's go back again and review our classic vertical differentiation model. You have consumers that have these vertical types theta. They have unit demands where your utility is theta times the quality of the good minus p. Let's assume that firms have fixed entry costs K. And then they have constant marginal cost c of s, where the higher quality products are more expensive to produce.

So consider a two-stage game. The first, the firms are simultaneously choosing whether to enter. And if they enter, they're going to choose which quality level to produce. And then, once they've done that, they compete in prices according to the vertical differentiation model.

And Shaked and Sutton have this classic paper that shows that you can get two different types of outcomes in this. First one is what they call the natural oligopoly case. So imagine that if I put s on this axis and I put-- what am I going to do?-- either the value that they get from s, theta s minus p, on this axis or I put c of s on this axis.

So imagine that-- imagine that you have relatively flat marginal cost curves. So imagine that marginal cost c of s looks like this, that there's just not that much cost to producing high-quality goods. And suppose that you have like theta lower bar times s looks like this. And theta upper bar s minus p looks like that.

So we have a model where everybody, even the lowest possible types, value quality more than it costs to produce the extra units of quality. And there's some range here, s lower bar, s upper bar of the qualities that are available to be produced.

In a model like this, what they call the natural oligopoly case, everybody should-- in an efficient market, everyone should be buying the highest possible quality. And so if you imagine what happens-- suppose you have lots of firms entering and producing different qualities. And then what are they going to set their prices at?

Well, if you have lots of firms in the market, the prices are going to fall-- these are prices; I'm putting dots in here-- the prices are going to have to fall towards cost as N goes to-- as the number of firms into this finite product space goes to infinity. But then if the prices fall towards cost, everyone is going to buy from the highest quality firm. And so you just can't have prices falling to cost, because if prices fall to cost, all these firms make zero sales, and then they earn negative profits after paying their fixed cost K.

So if you have this case where everybody should buy the same product, there's only a finite number of firms in the market. Again, this isn't an N goes to infinity as K goes to 0. As K goes to 0, you still get a finite number of firms. It may be this market can fit three firms or five or six or whatever, but there's a finite amount of entry that can ever occur if the same product is efficient for everyone to buy.

And then the second case they talk about is the specialization case. The specialization case would be if you have marginal costs that work like this. Marginal costs are initially low, but then they shoot up like that.

If the marginal costs look like this, what's efficient for the theta lower bar types? The theta lower bar types would pick the point-- what's efficient for them to do is where the marginal cost of producing extra quality is equal to the marginal benefit that they receive. So the theta lower bar types-- so this would be s star optimal for the theta lower bar types. They want to produce where-- they should be buying the good where the marginal value of increasing quality is exactly equal to their willingness to pay.

And the theta upper bar types-- maybe I should have made this steeper. [CHUCKLES] Let me make theta upper bar even steeper. The theta upper bar types are going to have a point where, again, they're going to buy up here, where it has this greater slope. So the s star theta upper bar, they should be buying this product. And everybody with a theta in between should be buying something in between. Everybody should be picking-- the optimal product for everybody is the point on that curve where it's an efficient quality for them.

So if you have a model like this where efficient quality-- everybody has a different efficient quality, then as N goes to infinity, you are going to get an infinitely divided model. So every product is efficient for someone. So if the number of customers is large, all those products get produced. They all sell to 1 over N of the people. They sell to the people whose prices are-- whose marginal values are close to their marginal-- whose optimal qualities are close to their quality and prices do fall towards marginal cost. But even though prices fall to marginal cost, everybody still buys it from a different firm.

So what happens with vertical differentiation is sometimes you can't get a lot of firms, but sometimes you get the specialization. And it all depends on whether different products are optimal for different people. Questions?

OK, so next thing-- next set of papers I want-- next paper I want to discuss is-- this paper is motivated by-- there's a large literature on firm dynamics and entry and exit that's generated by people looking at census data. So I don't know if I mentioned-- I think I mentioned this earlier in the semester.

So there's this great resource for IO economists which is called the Census of Manufacturers. So the Census of Manufacturers is conducted by the US Census Bureau every five years, all the years that end in twos and sevens.

Census of Manufacturers, it's a complete census of every manufacturing firm in the United States. In addition, the census does something called the Annual Survey of Manufacturers. It takes a subset of this and samples them to produce annual data instead of every-five-year data.

But anyway, they send out census forms to all manufacturing firms every five years, asking them lots and lots of questions. What do they produce? How much do they sell of every different, narrowly defined product? How many employees do you have? What's the value of your plant and equipment? How much electricity do you use?

So you get lots of data on these firms. Sometimes the accounting data of what are all their costs are hard to believe. But one thing that's very good in that data is, do you exist? What's your address? How many employees do you have? And so this lets you-- if you want to study entry and exit, you can just-- there's a windowless room at the NBER where one can access census data. Now, as a graduate student, one of the difficulties is that you try to access it through the NBER, and there's a two-year application process to be approved as a special sworn employee of the census.

And two years is a long time for a graduate student. I don't know if I should say this on camera, I think, but what people often do if they want to access census data is they get a RA job working with someone who has access to the census data, learn the census data well enough that they can describe all of the regressions and code that they will run when they get access to the census data, and then they have all the programs entirely written to just push the run button when their access is eventually approved two years later. And then they can write their paper using the Census of Manufacturers data.

But anyway, one thing that's very good to do with census data is you can get very good statistics on, how many firms enter every five years? How many firms grow? Of firms that enter in a five-year period, how many of them are still there five years later? Of the ones that are still there five years later, how many of them have grown? How many of them have shrunk? What's the relative profitability of the new firms versus the old firms?

So there's a set of old papers I used to teach here. Obviously, the classic one that's still on my syllabus is Dunne, Roberts, and Samuelson from the 1970s that first just wrote lots and lots of statistics on what entry and exit looked like in the census data. But you could search for things that have cited it to find many updates since then.

But let me say, some of the stylized facts here is that according to the census data, at least, there's a lot of entry and exit. And I think in some industries, we know this. If you watch restaurants or something like that, you're aware that new restaurants are born all the time, restaurants die all the time. There's a tremendous amount of entry and exit into the restaurant business.

Well, the surprising result from Dunne, Robertson, and Samuelson is that lots of other businesses seem to look like that too. Now, it's not the case that there's a tremendous number of new Apples and Samsungs every five years. But if you look at the firms that are classified into many narrowly defined businesses, it seems like many more are like the-- many more are like the restaurants than the Apple and Samsung phone producers.

Anyway, so what Jovanovic's paper is about, trying to discuss basic framework for how do we think about entry and exit in a way that would explain what's going on in the census data, when all these new firms are coming in, all these firms are exiting, and growth patterns, and so on.

So classic model Jovanovic is a macro-ish paper. So there's a lot of this assuming continuum of firms and whatever and assuming that everything works out evenly. So it looks a little different from I/O, but I wanted to cover it. So we start in this model-- imagine we have a continuum of small potential entrants that are out there. And these firms are going to have some fixed cost K of entering the industry.

And then if they enter the industry and then they decide to leave, they're going to have some liquidation value w that's less than K. So you're a restaurant, you enter, you pay K to set up your restaurant. When you're done, you can sell all your tables and your cooking equipment and the leases and whatever, all the money you spent on advertising and other things, that's sunk. But there's some fraction of-- there's some amount w that you can get back when you close up.

Jovanovic assumed that these firms have unknown types theta i. You can think of theta i as-- well, it's like the inverse of their productivity. So theta i is distributed normal 0 sigma theta squared. This is how good of a firm you are. You could think of it as, like in the restaurant case, everybody thinks they have food that people are going to like the taste of. But you never know until you open your restaurant and you see how popular it is.

So all these firms, they all have some d well, in fact, they all have-- in this model, they all have exactly the same prior. They all have the prior that my theta i is drawn from this distribution. But the only way to learn whether my theta i is low, in which I'm going to be a very profitable firm, or high, in which I'm going to be a very unprofitable firm, is to enter, start selling some-- open my restaurant, start selling to customers, see if people come and are willing to pay my prices, and I sell out or not.

So he does this with the thetas. Rather than quality of the food, he does this by your cost of producing as if it's your-- everyone's trying to come in. They're all going to try to produce some kind of steel. And then some of them end up being high-cost producers, some end up being low-cost producers, and that's what determines their profitability.

So firm i's period t cost, there's this some common function c of qit that applies to everyone. But then it's multiplied by f of theta i plus epsilon "it", where theta i is your permanent type and epsilon "it" is a shock. And what he's imagining is that f might be some function that looks like this.

And so we'll make this 1. So if my type is exactly 0, then-- or when I make this--

There are extra parameters in this model that probably shouldn't be there. So let's call this theta 0. So if my type is exactly theta 0, my cost is c of q. If my theta is negative infinity, then my cost is like 1/2 c of q. And if my theta is like positive infinity-- so I'm graphing f of x-- maybe my cost is twice c of q.

And so with these firms that are entering, if their costs are low, they're going to earn positive profits, high profits. If their costs are high, they're going to earn negative profits. But he adds this random shock in. So there's this random period-by-period shock. So I enter, I try to produce, this period I turn out that I had high costs. I don't know, am I a high theta type? Or was I just unlucky?

So again, if you think about it on the quality side, I enter, I start selling-- either this week, there's nobody at my restaurant; is it that my quality is low? Or did I just have a bad shock, a bad random draw that week?

And so firms are-- symmetric information model firms don't know their types. But every time they produce, they get a signal about what their type is. They update their beliefs. And then by updating their beliefs, they're going to change two things.

c is convex in this model. So if I think my cost is low, I'm going to produce a lot and continue going because I'm like, I can produce a lot and earn positive profits at the equilibrium level p star. If my cost is high, I'm only going to produce a little because I want to get the signal and find out if I'm good or not. But I'm only going to produce a little bit because I'm probably going to lose money on everything I produce.

So anyway, he treats the entrance as a continuum per period. And so with this continuum of entrance, you get a deterministic price sequence, p1, p2, p3. And he also assumes that the firms are all price takers, and they're just choosing q to maximize the expectation over theta i of the profits I'm going to get this period, which is pt, which I know, times q minus my cost of production, c of q, f of theta i plus epsilon "it".

And again, there's an expectation here because I don't know theta i. And actually, also there's going to be an epsilon "it" is also going to be random shock. So I guess this really should be expectation over theta i and epsilon "it".

So then what's the optimal behavior for a firm? Suppose that you think theta is very low. If you think theta is very low, then you think your costs are going to be very low. You know that the prices are. So you produce a very high quantity-- because you can produce a very high quantity with the increasing costs and still make a profit. So this is not a constant marginal cost model. This is a marginal cost-- this is c of q that's increasing in q.

If I have a medium expectation of theta, then I'm going to produce some low quantity. I'm going to hope that I'm better than I think I am, but I'm just going to produce the units I can produce cheaply, produce my initial cheap units, and then leave it at that and then hope that I find out I'm better, but at least make a profit on the ones that I produce.

If I have a high theta-- if I have a high expectation of theta, I think I'm a bad firm, but I'm uncertain about how bad I am, then I'm going to produce a low qit and stay in the market. And I'm just going to produce these very few units and try to sell those and hope that I'm better than I think I am. But if I'm high theta and I have low uncertainty, then I'm just going to take the liquidation value w and get out of the market.

OK, so firms are making this-- endogenously choosing both, What quantity do I produce? What quantity do I produce? And do I exit or not? And he separates these things where the amount I'll earn doesn't depend on the amount that I produced. And so if I produce, I'm going to produce the static optimal quantity. Yes.

AUDIENCE: So someone's belief or their expectation of theta i is just based off their previous observations of theta i?

GLENN ELLISON: It's based on their previous-- like, every period you get an observation of theta i plus epsilon "it". So you have a-- it's a normal learning model. I have a prior that it's normal 0 theta. These epsilons are also normal. So every time I enter, I get a normal signal, and then I just Bayesian update up and down. And the number of signals that I get determines whether my posterior is diffuse or narrow or tight.

So sometimes the older I am, the less uncertainty I have. Like, with normal distributions, it's just purely 1 over N. The uncertainty diminishes like 1 over N as I get N observations.

And so we notice that a basic model like this makes a number of predictions about firm dynamics that seem to fit with what the basic census data on firm dynamics look like. So some observations, small firms grow faster and fail more often. Why is that? Well, a lot of the small firms are-- a lot of the small firms are these people with high uncertainty.

The small firms are new firms that come in or firms that have come in and gotten one bad draw. They think their theta is high, but they're very uncertain about their theta. So they're producing small quantities. If they find out that they're better firms, they grow very quickly. These guys, the only way you can have a low expectation of theta is if you've gotten a lot of data and all the data says that your quantity is low-- your costs are low. Therefore, you're not going to grow very fast, because you've already got a lot of data. You're going to continue to think that your costs are low, and you're going to have roughly the same output.

So the small firms are firms who have a lot of uncertainty but think their costs are relatively high. Therefore, they grow fast if they stay in the market. But they also just exit and fail much more often. Bigger firms have higher profits. So profits are quite different across firms. The high-profit firms are the people who've been around for a while. They've learned that their theta is very low. Therefore, they can produce way out here on their cost curve and still make a profit. And they're making all these profits on the inframarginal units.

So the big firms are the-- big firms are the profitable firms, and they're profitable because they just are more efficient cost-wise than the other firms.

If you think about how our industry concentration and profits related-- and we have this sort of standard view that industry concentration and profits are related because we can put more firms into a Cournot model or more firms into a Hotelling circle model, markups drop. Here, this is like a perfect competition model.

But in equilibrium, industry concentration is correlated in the cross-section with high profits because an industry that happens to have a few very low-cost firms in it, those firms will earn very high profits. Whereas with another industry, all the firms have very similar costs, then you're just going to have a large number of firms, each earning individually low profits.

And so the firms that are going to look like high-profit industries, they're concentrated because they're the firms that just have a few firms that have theta draws much lower than everybody else's theta draw. And so you get this correlation between profits and concentration that comes from the distribution of realized costs.

So we get a profit concentration cross-section, but it's caused by a correlation with the omitted variable of distribution of firm costs rather than caused by competition and causal effect.

Industry concentration is also correlated with cross-sectional variance in profits. So again, if you have an industry that has-- if I do histograms of costs, if I have an industry that has a mass of firms that have very low costs and then some firms that have higher costs, this industry is going to end up very concentrated because the low-cost firms are going to produce a ton. All these firms are going to produce small amounts. And so we're going to see highly profitable firms and very nonprofitable firms.

If you're going to have an industry that's not concentrated-- an industry that's not concentrated is one that has much less cost heterogeneity. And so if it's got much less cost heterogeneity, no one produces a large quantity. And that also means that everyone's getting similar profits to everybody else because their costs are all similar to everybody else.

Note that these guys are all making essentially zero profits. And so it's like small numbers and 0, but small numbers and 0 are not so different in this model. So anyway, you get this. Concentration is both correlated with levels of profits and with cross-sectional variance in profits.

Anyway, so there's a large, large literature on the empirics of entry that bears out that many of these things seem to be true. And I guess this is the way we often think about-- we make sense of that data, is that it's not that all firms have the same technology. Some firms are just more efficient than others for no observable reason-- for no observable reason.

Any questions on this framework? I think that's all I had to say on it. Yep. So what I want to do in this lecture-- and I covered this-- I don't always teach Jovanovic. But I like to have some recent theoretical papers in the class. So anyway, this is a paper by Nikhil Vellodi, who's an assistant professor in Paris.

Anyway, so basically what Vellodi has done in this paper is gone back to Jovanovic 1982 and used Jovanovic 1982 to think about the effects of rating intermediaries and what types of policies we'd want rating intermediaries to have or what effects rating intermediaries have. So by rating intermediaries, he has-- thinking about firms like Amazon and Yelp, where on Amazon, you can leave feedback-- like Amazon, majority of the merchandise on Amazon is not actually sold by-- it's not by Amazon itself. It's sold by third parties selling on Amazon.

There are lots of-- you want to buy an umbrella on Amazon, you go there and you look at the reviews of that umbrella and whether people say that umbrella is good or bad, and then you buy it or not. You get Yelp or Tripadvisor or whomever that has lots of reviews of restaurants and tells you, Is this a good restaurant? Is this a bad restaurant? Consumers use that to learn about product quality.

And so there's this sort of challenge in coming up with these in this business model is that there are big firms that everybody buys from. On the big firms everybody buys from, there are going to be a zillion reviews and people are going to be able to find out the quality of those firms. But you have many, many small entrants and many, many new entrants all the time into these models. And it's the new entrants that people really want to know about and learn about whether their quality is good or bad.

But there's this effect of if you have new firms, people are hesitant to buy from them because they're new. And because people are hesitant to buy from them because they're new, there are very few reviews. And therefore there are very few reviews, therefore people can't learn about them. And so he wants to think about is, what would be the optimal design for a rating intermediary that wanted to promote an efficient level of competition?

In some sense, we think there could be-- why could there be too little entry in models? There could be too little entry here-- another reason I didn't put my theory in my first part was there'd be too little entry because people don't know the quality of new firms, and therefore they won't buy from new firms, and therefore no one will enter and have people learn about their quality.

Jovanovic was very convenient that the firms could enter and then produce these tiny quantities and get highly informative signals about their quality and know whether to start producing more. In practice, that's not something restaurants can do, is come in and produce a tiny quantity and get a lot of reviews on Yelp.

So anyway, models got a continuum of small potential entrants. So think of these as new restaurants or new guys just trying to sell black folding umbrellas on Amazon. They have a fixed cost K of entering. They're going to have some flow cost of operation. I didn't put a letter down here. Maybe this is c. But every period that I'm continuing to operate, I have to pay some flow cost. Just to make the model have a steady state equilibrium, he has some exogenous death rate delta. So every firm with some probability just disappears in each period. And then firms have some discount rate rho.

In his model, he's simplifying the theory by assuming that it's a discrete type model. All restaurants have either quality 0 or quality 1. So it's just a pure bimodal distribution. Again, like in Jovanovic, it's a symmetric information model. There's prior mean p0 that the firm has quality 1. And that information is symmetric among everyone in the market. No one knows or everyone knows new restaurants, there's probability p0 that their food is good, probability 1 minus p0 that their food is bad.

Every firm has exactly one table. And the firms can do is set prices qit for their table at time t. What he assumes is that you have this prior p0, and then you have a posterior-- every period, if somebody-- if you're operating, then reviews arrive according to a Poisson process.

And those reviews are going to make the quality, the posterior quality go up or down, whether it's a good review or a bad review. He goes with this continuous time, Brownian motion learning process where my posterior "pit" evolves according to dp "it" equals square root of lambda "it" "pit" 1 minus "pit" times dZit, where Zit is a Brownian motion.

But you can roughly think of this as lambda "it" is the review arrival rate. And each of these is like a normal signal. And so if it's a normal signal, I'm going to get information that either pushes me to the right or left. And the variance of-- since I have a Bernoulli quality, the variance of my distribution is just "pit" 1 minus "pit" if it's Bernoulli probability p.

So anyway, the new information is arriving, and I'm getting square root of lambda "it" is how much the standard deviation changes when you have lambda "it" reviews. And then dZit is a Brownian motion that's just positive or negative. So you can think of it as the limit as you get a number of normally distributed signals saying this one was good, this one was bad, this good, this bad, this good, this bad. That pushes you back and forth in like a Brownian motion process.

And why is he doing this in continuous time? He's doing this in continuous time instead of in discrete time because it's a state of the art thing, but it's also that sometimes continuous models are easier to solve than discrete models. You have to keep track of what your posterior could be. If you've had three and for and five and seven signals, the continuous model is easier to solve.

So every consumer in each period decides which-- so every consumer goes to exactly one restaurant. They decide which firm to visit. If excess consumers visited any type of firm, any firm, then service is rationed. There's a line outside, and only some of the people go in. Consumers get utility theta i plus epsilon ij if they're served by restaurant i.

So given that all the firm's priors are common knowledge, in equilibrium, everyone's going to get served with probability 1 at firms that have a rating of at least p tilde. And if your prior is-- like, you have one table, so it's going to be like-- you're going to clear the market. If your quality is high enough, you're going to price at some-- if your quality is at least p tilde, you're going to price at some equilibrium price that gets people to come to your restaurant. And you're going to price it exactly one customer comes.

If your price is attracting multiple customers, you should be raising your price and getting fewer customers until you fill your capacity. At all higher prices, everyone is going to be-- yeah, so you can tell Nikhil is a theorist not an IO economist. He's using q for prices and p for qualities, rather than p for prices and s for qualities.

So anyway, the firms are setting these prices q. In equilibrium, the prices are just p minus a constant. And that makes consumers indifferent to going to all high-quality restaurants and has all high-quality restaurants attracting exactly one customer.

Why you price it p minus w, it's because there are low-quality restaurants that people aren't going to want to go to that are willing to sell at a loss. Because they sell at a loss, they get people to come and review them. So in his model, basically the lambda "it", the review arrival rate, is going to be a sum of two terms.

It's going to be a sum. Lambda is going to be the number of customers who visit your restaurant because it's optimal to do so, plus epsilon. So there's this epsilon background rate of arrivals. We'll call them tourists from out of town who have no idea what the qualities are and just show up randomly at your restaurant. So most reviews come from the one person who knows your quality and buys there. But there is this epsilon arrival rate.

The firms that are hanging around that have low quality, they would like people to come to their restaurant and review it so they can learn if their quality is high and they can start charging more money for their restaurant if it's high or exit if their quality is low. But there's only so much they're willing to pay. So they're willing to price below cost to get people to come to their restaurant. But there's only so much you're willing to do that because, at some point, you're just better off exiting.

But anyway, so that's why you have this positive w. That is, everybody is giving consumers some surplus because the guys who are on the outside looking in are going to give the consumer some surplus to try to get the customers to come.

OK, so I'm going to write V of p for the value function of a firm that has posterior p. Free entry implies that V of p0 equals K. When the firms are entering, it must be that their lifetime expected discounted value of profits is exactly equal to the fixed costs. And then exit occurs when your posterior gets down to p lower bar, where V of p lower bar is 0. And V of p lower bar is 0 in this model without the scrappage because there's this flow cost of operation.

A complicated model, but I thought what I would do is here is just show you a couple of the graphs and talk through what ends up happening. So first observation-- so it's not-- yeah, so if you're a new firm and people-- in some sense, in equilibrium, if the number of customers is x, the x best restaurants are going to all get one customer each. And every restaurant below x is going to get no customers, other than the randomly arriving tourists.

But firms-- if you're just on the outside looking in-- so if we have this sort of posterior on my quality, I think it was p tilde I called it-- so everybody up here fills their restaurant. Everybody else here only gets the epsilon flow of tourists. If your quality is right here, you're willing to lose some money in order to get the customers to visit your restaurant to find out whether your restaurant is good so you can get in or out of the market.

So anyway, firms price somewhat below cost to attract customers. Because this guy is selling at-- this guy is willing to sell at a loss, everybody else has-- everybody else can only charge their marginal value above this guy-- their marginal quality above that guy. So consumers get-- consumers get some profit. What is the consumer surplus? The consumer surplus here is the amount of money this guy is willing to lose. Because the amount of money that this guy is willing to lose is-- the w is then whatever what you get at all higher-quality restaurants.

With full information, learning about new entrants is very slow. And firms exit after they just get some bad news. So here is his-- this is an example where p0, the prior, is in the range where you're not in the money on customers. So at a p0 restaurant, you're not getting very many-- you're out of the money, and customers are not coming to you.

And so what we get is, there's a lot of firms who just entered about who the prior quality is p0, and only epsilon people go to them. And so we get a lot of mass there because when there are only epsilon customers, people don't learn much, and your posterior about their quality doesn't move very fast. If you get one or two bad reviews, then you just exit.

So you start here, you get a bad review or two, you just exit the market. And then also, there aren't many firms down here with these qualities because two bad reviews, you leave. And then even if you would have gotten a third review coming back up, you're not there anymore.

Then there are fewer firms with qualities here because once your quality, your posterior gets above p upper bar, people learn about you very quickly. People learn about you very quickly, your distribution moves, and it either moves back to this range where no one's coming to your restaurant, or it moves up, and people learn it's good. And in this model, everyone is either 0 or 1. So very quickly, people learn that you're a quality 1 restaurant.

So we get this distribution of firm qualities where there are a lot of firms that are having trouble attracting anyone's reviews. They stay, hang out here, and their posteriors don't move much. And then once you get up to some threshold, you move very quickly, either back into the range where you don't learn much or into your quality is 1, and then people all start going to you.

Consumer welfare under any policy is the W that arises given the policy. So consumer optimality, like what's best for consumers if you're trying to-- if you're Yelp and you're trying to design my platform to be consumer optimal so that people will continue to use Yelp, what's optimal for the firm to do is to give firms the incentive to be willing to price below cost to have firms learn about them.

And so then what do you get about-- what would be the optimal policy if you're trying to get these guys to be willing to pay a lot to get more reviews? And it turns out that the answer is that you want to have a partially revealing customer policy where you don't publish all the reviews about high-quality firms.

What you want to do is, in some sense, as soon as a firm's posterior passes p tilde, you want to stop publishing any reviews about them whatsoever. And you do that because that way, by having a flat benefit from being quality p tilde or all the way up to here, when you have a flatter benefit from high quality, there's greater incentive to make the jump from here up to here.

And therefore, these firms are willing to price further below cost in order to make that jump. And then you are disadvantaging some firms. You're hurting firms up here, helping firms down here by not promoting-- providing quality information on them that lets them show that they're really good. But since new firms are much more likely to be here than they are to be here, when you're here, you're willing to pay more to get up. If you take some of the surplus away from the really successful firms and give it to the moderately successful firms, that raises the incentive to price in that way.

So what he argues is that what a platform would like to do in this model is to actually give incomplete information about restaurants and to selectively give information, where we restrict information about the good products so that we have more-- so that there's more of a benefit to having a pretty good reputation so that new firms will then invest more in trying to get the pretty good reputation. And then obviously, all the other firms are pricing off half the price to outcompete those guys.

OK, that is it. That's what I had for today. So Wednesday, I'm going to talk about empirical work on entry. I would say Bresnahan, Reiss is really the classic paper on entry. Again, Bresnahan, Reiss is a little old by now, so I would say that if you want to read things-- I will tell you what's in Bresnahan and Reiss with slightly more modern econometrics or what would be in it. And then maybe read Berry, Waldfogel, or look at one of the Bronnenberg, maybe Bronnenberg, Dhar, and Dubé or whatever. But I'll talk about those four papers.