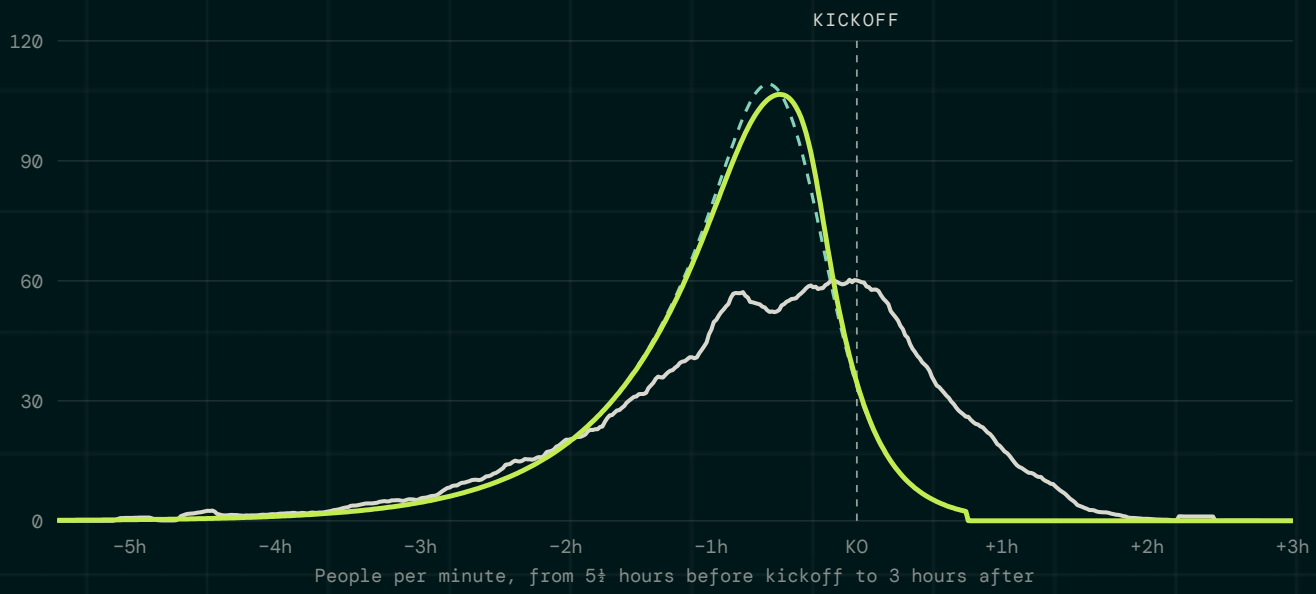
loft

PAPER

CROWD DYNAMICS

Modelling crowd arrival at stadium turnstiles

A choice model of arrival, and the gate capacity it implies.



PREPARED BY

James Katende and Mugalula Nickson

PUBLISHED

September 2026

READ ONLINE

loft.ug/research/arrival-model

ABOUT THIS PAPER

PREPARED BY	James Katende and Mugalula Nickson
ORGANISATION	Loft Technologies, September 2026
SERIES	Paper · Crowd dynamics
PUBLISHED	September 2026

CONTENTS

1	Introduction	3
2	Levels of description	4
3	Consequences for provisioning	7
4	Control	10
5	Numerical experiments	12
6	Scope and limits	12
7	Conclusions and research directions	13
	References	14

HOW TO CITE

Katende, J. and Mugalula, N. (2026). Modelling Crowd Arrival at Stadium Turnstiles. Paper. Loft Technologies.

Read online and download: loft.ug/research/arrival-model

14 → 10
Turnstiles a crowd of 9,000 needs, from all aiming at one moment to the calibrated spread

2.5
Times a minute late costs against a minute early (γ/β), set from the domain

0.43×
Queueing left when desired times spread out, against 0.56× from 20% more capacity

11×
The queueing when a gate falls short of the no-queue capacity by half

ABSTRACT

In this paper we present a model of spectator arrival at stadium turnstiles and derive from it the gate capacity at which no queue forms. We treat arrival as a choice: a supporter selects a time to reach the gate, trading the cost of waiting inside the ground against the cost of missing the start of the match and against the queue expected there, and under random utility the arrival density is the Gibbs measure over that cost. Its skew is governed by a single behavioural ratio, the cost of a minute of missed match against a minute spent waiting inside, which we set from the domain at $\gamma/\beta \approx 2.5$ rather than estimate. Because the queue is generated by the crowd's own choices the congested case is a fixed point, and solving it predicts that an undersized gate draws its crowd earlier and flattens its peak. The uncongested limit collapses to a closed form whose peak gives the capacity requirement, in quantities belonging to the crowd, not to the building, and convertible directly into a turnstile count: a crowd of 9,000 with a mean early lead of 42 min requires 14 turnstiles if its members all aim at the same moment, and 10 at the spread in desired times calibrated here. We then compare four quantities an operator can act on against a single objective, the person-minutes spent queueing, and find that dispersing desired times outperforms adding capacity. Nothing here is fitted: a turnstile records admission rather than arrival, and the two differ by the wait, so an admission log constrains the model only where the gate does not bind, and is least informative exactly where provisioning matters most. Turnout enters only as an exogenous scale, and the model makes no claim about who attends.

KEYWORDS

crowd arrival; turnstiles; gate capacity; departure-time choice; bottleneck model; random utility; queueing; stadium operations

§ 1 Introduction

This paper presents a model of spectator arrival at stadium turnstiles, and the gate provisioning that follows from it. Arrival is treated here as an outcome of choice, not as a draw from a distribution fixed in advance: each supporter weighs being early, being late and queueing, and the arrival curve of a crowd is what those choices aggregate to. This is the structure of the departure-time literature originating with Vickrey's bottleneck model[1], transposed to a setting in which the deadline is a kickoff and the bottleneck is a bank of turnstiles.

The crowd dynamics literature models what happens once people are inside a space and moving. The social force model treats each pedestrian as a particle under repulsive and attractive social forces[2]; continuum theories replace individuals with a density obeying a conservation law[3]; multiscale treatments connect the two[4]. The field is surveyed by Bellomo and Dogbé[5]. All of them take the population in the domain as given and resolve its motion. The question here comes before that: how many people present themselves at the boundary, and when. That is a problem of scheduling and not of locomotion, and the model below is built accordingly. The three scales at which § 2 describes it are nonetheless the standard ones of that field, and are used here in the same sense.

The problem is constrained by what a stadium can observe. A turnstile records the instant a credential is accepted, which is the moment a supporter reached the front of the queue rather than the moment they arrived at the ground. Writing A for admission time, R for arrival and W for the wait between them,

$$A = R + W, \qquad W \geq 0 \tag{1}$$

so an admission log gives A while the behaviour to be modelled is a statement about R. The two coincide only where $W \approx 0$, that is, where the gates are not the binding constraint. Where they do bind (the case that matters for provisioning) the record carries least information about the quantity the model needs, and a tail of admissions continuing past kickoff is consistent both with a crowd that arrived late and with a crowd that arrived on time and queued. Calibration is correspondingly difficult. The queue is never observed directly, and behavioural parameters such as the relative cost of a minute of lateness against a minute of earliness cannot be recovered from a log the gate has flattened.

We proceed as follows. § 2 sets out the model at three scales of observation, closing it through the queue so that the congested case is a fixed point; § 2.6 gives the closed form to which the uncongested limit collapses. § 3 derives what follows for provisioning, in particular the capacity at which no queue forms and its conversion into a turnstile count. § 4 compares the quantities an operator can act on against a single objective, the person-minutes spent queueing. § 5 reports numerical experiments over the model's parameters. § 6 states the assumptions the results rest on and the measurements that would test them.

§ 2 Levels of description

2.1 Preliminary notions

A crowd converging on a stadium can be described at several scales, and the choice of scale changes the modelling framework considerably. The finest scale follows one supporter deciding when to set out, weighing the cost of being early against the cost of being late and against the queue expected at the gate. An intermediate scale discards the individual and keeps a density over arrival times, obtained by treating that decision as imperfect and aggregating over the crowd. The coarsest keeps only the flow through the gate, which is the scale at which a venue is provisioned and the only one a turnstile records. We use all three, in that order: § 2.2 gives the cost a supporter faces, § 2.3 the density it induces, § 2.4 the flow a gate of finite capacity produces from that density.

The scales are coupled in both directions. The gate enters the description twice: once as the mechanism that converts arrivals into admissions, and again inside the cost that governs arrival, since a supporter who anticipates a queue sets out earlier. The second coupling is what makes the congested case a fixed point rather than a composition of three independent layers. It is also the one an admission log cannot see, since the log is a record of the first.

The size of the crowd is exogenous throughout. Whether a supporter attends at all is settled by competition, opponent and price well before the gates open, and nothing here bears on it: every quantity below is a share of N, the number who come.

2.2 Microscopic: a single supporter

Write t for arrival time relative to kickoff and t^* for the time a supporter would ideally be inside. We make four assumptions.

(A1) Schedule delay is linear and asymmetric. A minute spent waiting inside the ground costs β; a minute of the match missed costs γ; a minute in the queue costs α. This is the standard cost function used in the departure-time literature[6]. Its substantive content is $\gamma > \beta$: a minute of the match missed costs more than a minute spent waiting inside.

$$C(t) \;=\; \underbrace{\beta\,(t^{*}-t)^{+}}_{\text{early}} \;+\; \underbrace{\gamma\,(t-t^{*})^{+}}_{\text{late}} \;+\; \underbrace{\alpha\,W(t)}_{\text{queueing}} \tag{2}$$

2.3 Mean-field: from one supporter to a density

(A2) Choice is imperfect. Supporters do not solve this problem exactly. They have incomplete information about the queue and reasons of their own for preferring one time to another. We treat this with random utility and Gumbel errors[7], the standard discrete-choice assumption, under which the probability of choosing an arrival time is a Gibbs measure over its cost. The parameter θ sets how sharply the crowd optimises.

$$p(t) \;=\; \frac{1}{Z}\exp\!\left(-\frac{C(t)}{\theta}\right), \qquad Z=\int \exp\!\left(-\frac{C(u)}{\theta}\right)\mathrm{d}u \tag{3}$$

2.4 Macroscopic: the gate

(A3) The queue is endogenous. W is not exogenous. With N supporters and a gate of capacity s,

$$W(t)=\frac{Q(t)}{s}, \qquad \dot{Q}(t)=N\,p(t)-s\,\mathbf{1}_{\{Q>0\}} \tag{4}$$

which closes the system: arrivals determine the queue and the queue determines arrivals. The gate is therefore inside the model, not a filter applied to its output. This also yields a prediction that a fixed arrival distribution cannot express. At a congested gate the crowd *shifts earlier*, since αW penalises arriving into a queue, so arrival is not independent of throughput.

(A4) Desired times are not identical.[8] Supporters do not agree on the minute they wish to be inside: some come for the warm-up, some for the anthem. Taking $t^{*}\sim N(m,\sigma_t^{2})$ and integrating over it leaves the model in closed form, as a sum of two exponentially-modified Gaussians:

$$p(t)=\frac{ab}{a+b}\Big[e^{\,ax+\frac{a^{2}\sigma_t^{2}}{2}}\Phi\!\Big(-\tfrac{x}{\sigma_t}-a\sigma_t\Big)+e^{-bx+\frac{b^{2}\sigma_t^{2}}{2}}\Phi\!\Big(\tfrac{x}{\sigma_t}-b\sigma_t\Big)\Big], \quad x=t-m \tag{5}$$

with $\sigma_t\to 0$ returning (7) exactly. Two consequences follow. First, the density becomes differentiable at the peak, where a common desired time produces a corner. Second, the aggregate peak moves earlier than the typical desired time: at $\sigma_t=12$ min the crowd peaks at T−36 while the average supporter aims at T−30. This is the usual effect of convolving an asymmetric density, whose mode moves towards the heavier side.

2.5 Several gates

A venue admits through several gates, and supporters choose one by proximity or by ticket block, not by queue length, so uneven use across a bank of turnstiles is the normal case[9]. Write c_i for the capacity of gate i and w_i for the share of the crowd arriving at it. Gate i saturates once total demand exceeds c_i/w_i, so the venue saturates at the smallest such value:

$$s_{\text{eff}}=\min_i\frac{c_i}{w_i} \;=\; \frac{1}{k}\sum_i c_i, \qquad k=n\max_i w_i \;\geq\; 1 \tag{6}$$

Effective throughput follows the worst-loaded gate rather than the total installed capacity. Write k for the imbalance: the largest ratio, across gates, of a gate's share of the crowd to its share of the installed capacity. Where the gates are equally provisioned this is $k = n\ \max_i w_i$ for an n-gate venue, and where they are not, the capacity share is used directly. The venue delivers $1/k$ of what it has built, so balance acts as a capacity multiplier. A venue in this state can queue at every gate while turnstiles stand idle at some of them.

2.6 The uncongested limit

Set $W \equiv 0$. The exponential of a piecewise-linear cost is piecewise exponential, and normalising gives a closed form in two rates, $a = \beta/\theta$ and $b = \gamma/\theta$:

$$p(t) \;=\; \frac{ab}{a+b}\,\exp\Big(-a\,(t^{*}-t)^{+} - b\,(t-t^{*})^{+}\Big), \qquad a=\frac{\beta}{\theta},\; b=\frac{\gamma}{\theta} \tag{7}$$

Three properties follow immediately, and none of them was put in by hand.

$$\begin{aligned} \text{mode} &= t^{*} && \text{exactly} \\ \text{median} - \text{mode} &= -\tfrac{1}{a}\ln\tfrac{2b}{a+b} && = -15\ \text{min} \\ \Pr(t>0) &= \tfrac{a}{a+b}\,e^{\,b\,t^{*}} && = 5.0\% \end{aligned} \tag{8}$$

The mode is t^{*} exactly, so the crowd peaks when it wants to be inside and no separate parameter is needed. The density at kickoff is finite and near its maximum, so the model requires no correction at the whistle. The post-kickoff tail is the same distribution continuing; no second one is appended.

The third expression in (8) is the untruncated tail. Arrivals stop when the gates shut, and truncating at c reduces the post-kickoff share from 5.0% to 4.6%. The closed expressions here are quoted untruncated because they are otherwise unreadable; § 5 evaluates the truncated form, and reports a slightly larger figure again whenever the gate is congested, for the reason given below.

The asymmetry of the density reduces to a single ratio:

$$\frac{p(t^{*}-\tau)}{p(t^{*}+\tau)} \;=\; e^{(b-a)\tau} \;=\; e^{(\gamma-\beta)\tau/\theta} \tag{9}$$

A supporter is $e^{(\gamma-\beta)\tau/\theta}$ times as likely to be τ minutes early as τ minutes late. The long thin lead-in that every ingress curve shows is therefore a consequence of $\gamma > \beta$, and does not have to be imposed by selecting a skewed distribution.

Calibration. Two domain facts fix the two rates: crowds peak about 30 min before kickoff, and the median supporter is inside about 15 min before that peak. Solving the landmark equations (8) gives

$$\frac{\gamma}{\beta} = \frac{b}{a} = 2.5, \qquad \frac{1}{a} = 42\ \text{min}, \qquad t^{*} = T - 30 \tag{10}$$

so lateness is priced at about 2.5 times earliness per minute. That figure can be argued with: the departure-time literature places the equivalent commuting ratio[6] near three to four, and the assertion that a football crowd has less at stake than a commuter is a claim about the world and not about arithmetic.

WHY THE PARAMETERS ARE NOT ESTIMATED.

We set them from the domain instead of fitting them, because an admission log cannot discipline them: where the gate binds, the log measures the gate, and scoring an arrival model against such a log rewards a flatter curve than the crowd produced. Deriving the curve also has the advantage that γ/β is a behavioural quantity with published values in other settings, which the width of a fitted distribution is not.

Admitted flow. The model above is *demand*: presentations per unit time. What a log records is demand passed through a finite gate. With capacity C per minute and a queue Q carried from one minute to the next, admissions in minute k are

$$a_k = C\left(1 - e^{-(Q_{k-1}+d_k)/C}\right), \qquad Q_k = Q_{k-1} + d_k - a_k \tag{11}$$

The throughput mechanism has two cases. When C exceeds peak demand, the admission curve equals the arrival curve and the log is an unbiased estimate of behaviour. When it does not, the log measures C, and the excess demand is displaced to the right, appearing as a tail of admissions past kickoff.

Table 1 collects the model's parameters. Two are fixed by the domain facts above, one is derived from them, and the remainder are either stipulated or exogenous; the distinction matters because § 6 traces which results survive if each is wrong.

Table 1. Parameters of the model, their values, and where each comes from.

SYMBOL	QUANTITY	VALUE	ORIGIN
t^*	Desired time inside the ground	KO − 30 min	Domain: crowds peak about 30 min before kickoff
$L = 1/a$	Mean early lead	42 min	Domain: median inside about 15 min before the peak
$r = \gamma/\beta$	Lateness aversion	2.5	Derived from the two above
α/β	Queue aversion	2.0	Standard ordering; not identified by an admission log
σ_t	Spread of desired times	12 min	Stipulated; nothing observed here constrains it
θ	Scale of choice noise	—	Only the ratios β/θ and γ/θ are identified
N	Crowd that comes	9,000	Exogenous; the model makes no claim about turnout
C	Gate capacity	varies	The quantity being provisioned
c	Gate closing time	KO + 45 min	Operational; truncates the tail
k	Gate imbalance	1.00	Balanced unless stated; a property of the venue, not of the crowd

§ 3 Consequences for provisioning

This section derives what the postulates of § 2 imply for provisioning: the shape of the arrival density, the number of free parameters it carries, the capacity at which no queue forms, and the two ways a gate below that capacity fails.

From (A1): exponential tails, and a test of linearity

A cost linear on each side of t^* produces a density that is exactly exponential on each side, so the log-density is piecewise linear with slopes $+a$ and $-b$. Demand halves every $\ln 2/a$ minutes before the mode and every $\ln 2/b$ minutes after it:

$$\underbrace{\tfrac{\ln 2}{a} = 29 \text{ min}}_{\text{demand halves, before } t^*}, \qquad \underbrace{\tfrac{\ln 2}{b} = 12 \text{ min},}_{\text{after } t^*} \qquad \frac{\ln 2/a}{\ln 2/b} = \frac{\gamma}{\beta} \tag{12}$$

The ratio of the two half-lives is r, so (A1) can be tested without estimating either rate. An uncongested log plotted on logarithmic axes should give two straight lines meeting at the mode. Curvature on either side would indicate a schedule-delay cost that is not linear (convex, if supporters weigh the first minutes of lateness disproportionately), and no choice of a or b would recover it.

From (A2): the number of free parameters, and the role of θ

Gumbel errors give a Gibbs measure in which β, γ, θ appear only as the ratios β/θ and γ/θ. The model therefore has two free parameters rather than four. No admission data can separate a patient crowd from a noisy one, since only the ratio is identified.

Since $L = \theta/\beta$, the capacity requirement derived below scales as $1/\theta$:

$$s^* = \frac{N\beta}{\theta} \cdot \frac{r}{1+r} \xrightarrow[\theta \to 0]{} \infty \tag{13}$$

A crowd that optimised perfectly would arrive together at t^* and require unbounded capacity. Halving θ doubles the turnstiles required, from 14 to 28 for a crowd of 9,000; a tenfold reduction requires 139. Under (A2) the dispersion that makes a gate feasible is attributable to imperfect optimisation and not to heterogeneous preferences. The result holds even for a crowd that agrees exactly on t^*.

From (A3): the capacity at which no queue ever forms

The closed form (7) peaks at t^*, and its height there is $ab/(a+b)$. A gate that can serve that rate never accumulates anybody. Writing $L = 1/a$ for the mean early lead and $r = \gamma/\beta$:

$$s^* \;=\; N\,\frac{ab}{a+b} \;=\; \frac{N}{L} \cdot \frac{r}{1+r}, \qquad L = \frac{1}{a},\; r = \frac{\gamma}{\beta} \tag{14}$$

Every term in this expression is a property of the crowd. The requirement is set by how the crowd distributes itself in front of the gate, not by the stadium's holding capacity. At $L = 42$ min and $r = 2.5$, a crowd of 9,000 needs 153 admissions per minute (14 turnstiles at the standard rate of 660/h), and a crowd of 20,000 needs 31 turnstiles.

Expression (14) is derived in the homogeneous limit of § 2.6, where every supporter shares one desired time. Under (A4) they do not, and integrating over t^* lowers the peak the gate must serve, so it lowers the requirement with it. At the calibrated $\sigma_t = 12$ min the same crowd of 9,000 needs 113 per minute instead of 153, which is 10 turnstiles instead of 14; at $\sigma_t = 30$ it is 78. The closed form is therefore an upper bound on the requirement, and the requirement falls as the spread of desired times widens. § 5 reports both: the reported s^* tracks the calibrated spread, and setting the spread to zero returns 153.

This is not the same calculation as the Green Guide's[10]. That Guide caps the rate of entry at 660 persons per entry point per hour and takes entry capacity to be that rate multiplied by the number of entry points, which is an *hourly* figure; s^* is a requirement on the *peak minute*. Setting the two equal gives the condition under which the Guide's rate suffices:

$$\underbrace{11n}_{\text{Guide, per minute}} \geq \underbrace{\frac{N}{L}\cdot\frac{r}{1+r}}_{s^*} \quad\Longleftrightarrow\quad L \geq \frac{N}{n}\cdot\frac{r}{11(1+r)} \tag{15}$$

A ground filled to the Guide's own entry capacity of 660 per point requires a mean early lead of 42.9 min. The value calibrated in § 2 from the mode and median of football crowds, without reference to the Guide, is 42. The Guide derives 660 from measured rates of entry, not from a model of arrival[10]. On this reading it encodes an arrival profile with a lead close to 43 min, which is why the two methods agree there.

Applied to the Guide's own worked example for a football ground (15 entry points, a final capacity of 5,730), the stated entry capacity is 165 per minute against a peak requirement of 97, and entry is not binding at any stand. The methods do not conflict on that ground. They diverge for a crowd arriving more compressed than $L \approx 43$ min, and under the reduced rates the Guide itself anticipates: its contingency for enhanced searching puts the rate at 350 per point per hour, which for the same ground is 87.5 per minute against the same requirement of 97, so the ground becomes entry-bound under a measure applied for safety.

The Guide also requires that where the recorded rate of entry falls below 660 for reasons other than low demand, the lower recorded figure is the one used, and that rates be measured annually. The imbalance of § 2.5 is one such reason, and it is measurable from records the operator already holds.

The same expression (14) prices the alternative to building more gates. Capacity falls with L: a crowd persuaded to spread itself over 60 min instead of the calibrated 42 requires 107 per minute instead of 153, which is 4 fewer turnstiles for the same people. Under this model staggered or timed entry substitutes directly for built capacity.

The cost of undersizing, and the point at which the gate stops coping

Below s^* a queue forms, and the burden grows far faster than the shortfall. Table 2 solves the fixed point across capacities, measuring the burden as person-minutes spent queueing and expressing capacity as a multiple of the homogeneous s^* of 153 per minute, so that it and the requirement above share one reference point.

Table 2. Queueing burden and peak wait against capacity, as multiples of the homogeneous s^* of 153/min.

CAPACITY	QUEUEING BURDEN	PEAK WAIT
1.00 × s^*	1.0 ×	—
0.78 × s^*	1.9 ×	3 min
0.65 × s^*	3.6 ×	6 min
0.46 × s^*	10.7 ×	16 min
0.29 × s^*	31.1 ×	39 min

A shortfall of a fifth roughly doubles the burden; a shortfall of a half multiplies it elevenfold. We deliberately do not quote a single exponent: fitting a power law across this range gives about $s^{-2.8}$, but the local slope is not constant: nearer 1.8 just below s^*, above 3 in the middle, and falling again at the bottom. The behaviour is a crossover, not a scaling law, and a single power would misrepresent it.

A second and qualitatively different failure appears further down. So long as the gate can clear the crowd before it shuts, everyone is admitted late rather than refused. Below a threshold it cannot, and the residue is locked out:

$$\text{exclusion} \iff \int_{-\infty}^{c} \min\big(Np(t),\, s\big)\,\mathrm{d}t \;<\; N, \qquad s_{\text{crit}} = \kappa(L, r, c)\, s^*, \quad \kappa \approx 0.15\text{–}0.6 \tag{16}$$

At the calibrated parameters the threshold is 0.28 × s*. It is invariant to crowd size, taking the value 0.28 at 4,000, 9,000 and 20,000, and invariant to how early the analysis window opens. It varies with the other parameters, rising to 0.39 for a crowd spread over 60 min and falls to 0.15 for one arriving inside 25 min, and it moves with how long the gates stay open after kickoff: 0.18 at $c = 90$, 0.58 at $c = 15$. The range across these settings is therefore about a sixth to three fifths of s^*, and the point estimate should not be used on its own.

The symmetric case is degenerate: at $\gamma = \beta$ the threshold exceeds s^* itself, because a crowd with no aversion to lateness puts mass beyond the closing time and some supporters are excluded at *any* capacity. Lateness aversion is therefore a necessary condition for a capacity solution to exist.

Above the threshold an undersized gate produces delay; below it, exclusion. Nothing in an admission log separates the two regimes: a gate at 0.30 and a gate at 0.26 differ in whether some supporters never got in, and neither records the people it turned away.

None of these results is measured. Each follows from a postulate and fails if that postulate is wrong: curvature in a log-density refutes (A1), any evidence that patience and optimisation are separately identified refutes (A2), and a queue that grows in proportion to the shortfall rather than as a power of it refutes (A3). Each names the measurement that would decide it.

§ 4 Control

The model contains four quantities an operator can act on: the capacity of the gate, the spread of desired times, the time at which the gates open, and the time at which they shut. We compare them on one objective, the person-minutes spent queueing, holding the crowd at 9,000 and the gate at 100 per minute, which is below the requirement of § 3 and therefore congested. Table 3 reports the comparison.

Table 3. Four interventions compared on queueing burden and peak demand, crowd 9,000, gate 100/min.

INTERVENTION	QUEUEING BURDEN	PEAK DEMAND
None	1.00 ×	106/min
Capacity raised 20%	0.56 ×	109/min
Desired times spread, σ_t 12 → 30 min	0.43 ×	78/min
Gates open earlier, L 42 → 60 min	0.45 ×	86/min
Desired time moved to KO − 60	0.99 ×	106/min
Gates close 45 min later	0.98 ×	105/min

Dispersal outperforms capacity

Widening the spread of desired times reduces the burden to 0.43 of its original value. Raising capacity by a fifth reduces it to 0.56. Dispersal acts on the peak, which falls from 106 to 78 per minute, while capacity leaves the peak untouched and serves it faster. Since the requirement s^* is proportional to the peak, dispersal lowers the target as well as the load, and the two effects compound.

Timed entry, staggered turnstile allocation by ticket block, and any pre-match attraction that appeals to part of the crowd instead of all of it act on σ_t. Opening the gates earlier acts on L and performs almost as well at 0.45.

Two interventions that do not reduce the queue

Moving the desired time from 30 to 60 min before kickoff leaves the burden at 0.99 and the peak unchanged. The distribution translates without changing shape, so the gate meets the same maximum 60 min earlier. Moving when a crowd wants to arrive does not help unless the spread of those times widens with it.

Extending the closing time from 45 to 90 min after kickoff leaves the burden at 0.98. The queue forms before kickoff and is exhausted well before the original closing time, so the extension applies to a period in which no queue exists. It affects the exclusion threshold of § 3, which is a different failure, and leaves the queue unchanged.

Both are interventions an operator might reach for first, and under this model neither reduces the queueing burden. The two that do work act on the width of the arrival peak; these two move it or accommodate it, and the peak the gate must serve is unchanged.

Balance is a fifth quantity

§ 2.5 gives a venue's effective capacity as its installed capacity divided by k, so rebalancing the gates acts on the same term as building more of them, and a venue at $k = 2$ recovers as much by evening its loads as by doubling its turnstiles. It is absent from Table 3 because it is not a change to the crowd or to the timetable but to which gate a given supporter uses, and it is bounded: no amount of balancing takes a venue past the capacity it has installed. § 5 applies it directly, dividing installed capacity by k before the queue sees it. The quantities it needs, admissions per gate against turnstiles per gate, are already in the records a venue keeps.

§ 5 Numerical experiments

This section evaluates the equations of § 2 interactively. The two solid curves are both admissions, one synthetic and one modelled, so the eye compares like with like. The dashed curve is the model's demand, meaning the arrival that would occur if the gate never queued; it is drawn faint because it is not the quantity a log records (§ 1). Reducing the capacity separates the modelled admissions from it.

The reference curve is synthetic in the sense that it belongs to no single fixture: several recorded fills are each read as a rate, each normalised to its own crowd so that a large one cannot dominate, and the profiles are then combined. It is drawn as a shape to compare against and not as evidence, and nothing in this paper is inferred from it. What it shows is that a fill is late-heavy where the model's demand is not: it peaks near the whistle and continues past it, because a gate that is the binding constraint admits at its ceiling straight through kickoff.

Try the model: change the crowd, the gate, the clock

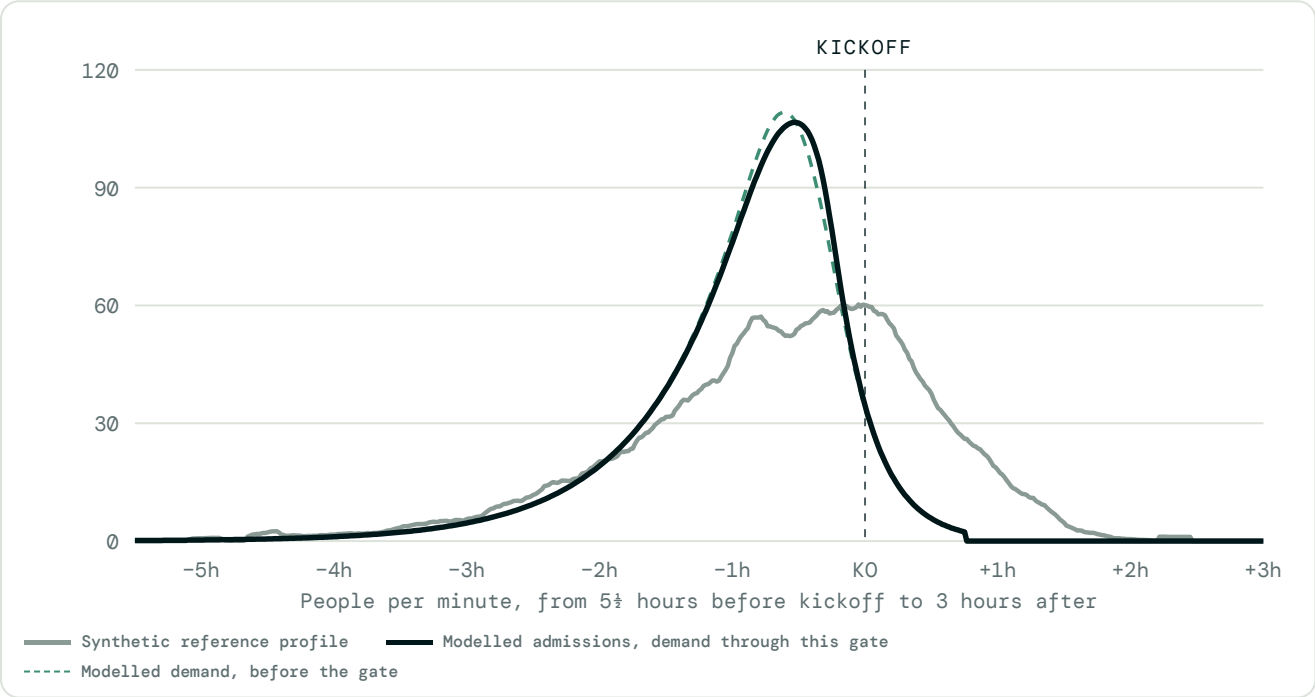


Figure 1. The model evaluated at the settings above. The dashed curve is demand, and the solid curves are admissions through the modelled gate and the synthetic reference profile. Lowering the capacity separates demand from admissions, which is the mechanism of § 2.4 seen directly.

§ 6 Scope and limits

The model rests on assumptions (A1)–(A4) of § 2, and its parameterisation comes from the domain. Nothing here has been fitted to a set of observations, so nothing here has been corroborated by one either: what follows is a statement of what the model asserts and of what would falsify it.

W is never observed. Attributing admissions after kickoff to queueing instead of to late arrival is an inference from (A3), and a measurement of queue length at the turnstile would settle it directly.

The queue of § 2 holds C fixed. A real gate does not open at its full rate: staffing, searching and the arrival of stewards make capacity itself a function of time, and while a gate is the binding constraint its admission rate *is* its capacity, so an admission log from a saturated gate traces that function rather than the crowd. Modelling it

would require a second time-varying quantity, which we do not attempt here. The consequence for § 3 is that a capacity ramping up towards kickoff meets the peak later than a constant one, so s^* understates what a gate that opens slowly must eventually deliver.

(A4) integrates over t^* with a Gaussian, chosen for tractability. Nothing here constrains the shape of that spread, and a skewed or bimodal one (a crowd divided between those who come for the warm-up and those who arrive for the whistle) would give a different peak. Its magnitude is equally unconstrained, and $\sigma_t = 12$ min is a stipulation.

(A4) allows supporters to disagree about t^*, but they disagree around a single event. The model does not describe a venue whose crowd is drawn by several staggered attractions, where the spread of desired times would be multi-modal instead of dispersed about one value, nor an event with no fixed start that people drift into across an evening.

We set $\alpha/\beta = 2$ on the standard ordering in the departure-time literature. The consequences in § 3 follow from that ordering: the capacity requirement, the superlinear cost of undersizing, and the threshold below which supporters are excluded rather than delayed. If the ordering is wrong, those consequences are wrong with it, and an admission log cannot settle it either way, since the log is flattened by the very gate whose effect would have to be separated out.

Departures are neither modelled nor recorded, so every quantity here is cumulative entry. This does bound occupancy from above: since nobody leaves before the match starts, cumulative admissions are exact occupancy up to kickoff and an upper bound thereafter. The bound is adequate for the operational question, which is how many supporters are still outside when the whistle goes, and that is answered by entries alone. Occupancy after kickoff, and with it egress planning, requires a record of departures, which turnstile admission logs do not contain.

§ 7 Conclusions and research directions

An admission log records the gate as much as the crowd, and the two cannot be separated by inspection. We have set out a model in which arrival time is the outcome of a choice, described it at three scales, and shown that the resulting density carries its asymmetry in a single behavioural parameter. Because the queue is generated by the crowd's own choices, the congested case is a fixed point rather than a composition, and solving it predicts that an undersized gate draws its crowd earlier and flattens its own peak.

The capacity rule of § 3 is the most readily transferable result. It is a closed form in quantities belonging to the crowd, it converts directly into turnstile counts, and it is falsifiable in the field: a venue that sizes its gates by it and still queues has refuted (A1).

Three measurements would each settle something the model currently assumes. Queue length at the turnstile, or arrival at the perimeter instead of at the gate, would identify α and with it the anticipation effect the closed loop predicts. Fixtures of comparable standing whose gates differ in capacity would test whether an undersized gate really does produce an earlier crowd. And a record of how many credentials reached supporters, rather than how many were generated, would make attendance measurable, closing the one quantity the model must take as exogenous.

None of these is available here, and the model is offered accordingly: as a set of consequences that follow from four stated assumptions, each of which names the measurement that would refute it. An arrival curve fitted to admission records would look better supported and would be worth less, because the gate flattens the record precisely when the gate is what matters.

References

1 W. S. Vickrey, “Congestion theory and transport investment”, *American Economic Review* 59(2), 1969, 251–260. The bottleneck model: arrival time chosen against a deadline under a queue the arrivals themselves create, which is the structure (A1)–(A3) follow.

2 D. Helbing and P. Molnár, “Social force model for pedestrian dynamics”, *Physical Review E* 51, 1995, 4282–4286. The microscopic standard for pedestrian motion, and the scale at which crowd models usually begin.

3 R. L. Hughes, “A continuum theory for the flow of pedestrians”, *Transportation Research Part B* 36, 2002, 507–535. The macroscopic counterpart, in which individuals are replaced by a density obeying a conservation law.

4 E. Cristiani, B. Piccoli and A. Tosin, *Multiscale Modeling of Pedestrian Dynamics*, Springer, 2014. Connects the two, and is the source of the microscopic–mesoscopic–macroscopic ordering § 2 follows.

5 N. Bellomo and C. Dogbé, “On the modeling of traffic and crowds: a survey of models, speculations, and perspectives”, *SIAM Review* 53, 2011, 409–463. A survey of the field this paper sits beside without entering: it models motion within a space, where the question here is arrival at its boundary.

6 K. A. Small, “The scheduling of consumer activities: work trips”, *American Economic Review* 72(3), 1982, 467–479. Source of the linear asymmetric schedule-delay cost of (A1) and of the estimated lateness-to-earliness ratios referred to in § 2, which fall near three to four for commuting.

7 D. McFadden, “Conditional logit analysis of qualitative choice behavior”, in P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press, 1974, 105–142. Random utility with Gumbel errors, giving the Gibbs measure of (A2) and the scale parameter θ.

8 R. Arnott, A. de Palma and R. Lindsey, “Economics of a bottleneck”, *Journal of Urban Economics* 27(1), 1990, 111–130. Equilibrium arrival curves for the bottleneck with heterogeneous desired times, the case (A4) treats by integration.

9 G. K. Still, *Crowd Dynamics*, PhD thesis, University of Warwick, 2000, ch. 3: on the limitations of estimating ingress capacity from turnstile counts, where averaging conceals surges.

10 Sports Grounds Safety Authority, *Guide to Safety at Sports Grounds* (the Green Guide), 6th ed., 2018, sections 10.9–10.11: the rate of entry is not to exceed 660 persons per turnstile or other entry point per hour, and entry capacity is that rate multiplied by the number of entry points.