Net Neutrality and Inflation of Traffic

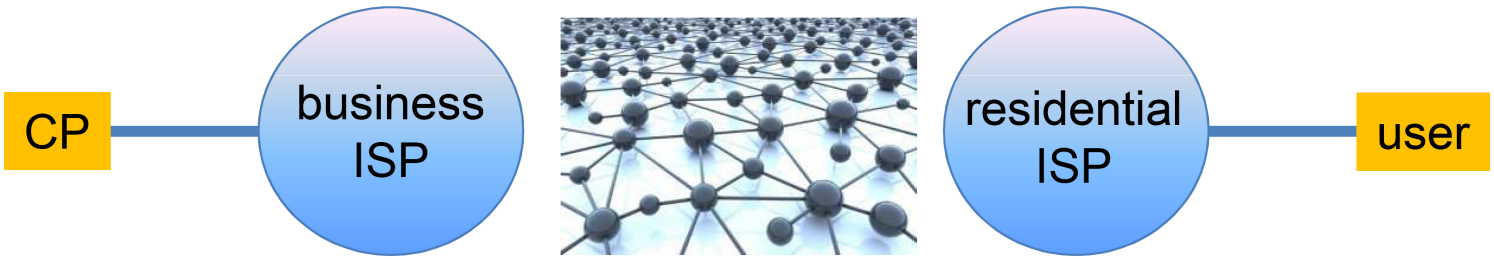
Martin Peitz (MaCCI, University of Mannheim and CERRE)
Florian Schuett (TILEC, CentER, Tilburg University)

Symposium in Honor of Jean Tirole
The Hague, 9 December 2014

The net neutrality debate (1)

A stylized depiction of the internet

- Traffic delivered through the backbone, best-effort principle ("all packets created equal")



Internet backbone

- Access charges:
  - Content Provider (CP) pays business ISP
  - User pays residential ISP
- No payment between CP and residential ISP
- No differential treatment of packets based on:
  - Type of content
  - Source of content (sender)

The net neutrality debate (2)

Key questions

- Should residential ISPs be allowed to charge CPs?
  - Sometimes considered "double charging"
  - But common in two-sided markets, and monopolist platform chooses right structure of prices
- Should residential ISPs be allowed to offer different tiers of service?
  - Fast lane and slow lane
  - 2nd-degree price discrimination
- Should residential ISPs be allowed to prioritize certain content?
  - Deep packet inspection (may raise privacy concerns)
  - Competition concerns when ISPs are vertically integrated (foreclosure), but not obvious why internet requires special treatment

Net neutrality and congestion

- Deviations from net neutrality (NN) have run into strong opposition
- But often there are efficiency justifications
- In this paper, want to draw attention to the responses by CPs
- Facts:
  1. There is congestion (limited bandwidth), leading to delay
  2. Some types of content are more sensitive to delay than others (e.g., Skype call vs email)
  3. Certain techniques used to minimize delay change the volume of traffic (base model: more traffic)
- Externalities in traffic generation: individual actions by CPs affect aggregate congestion on the network (common property resource)

Congestion control techniques

- More traffic:
  - Forward-error-correction (FEC) schemes may be used to protect video packets (Skype reported to react to persistent packet losses by increasing packet size; de Cicco, Mascolo, Palmisano, 2011)
  - Multiple multicast trees to provide redundant paths
- Less traffic:
  - Compression technology
  - Quality reduction (e.g., lower video resolution)
  - Reducing advertising intensity or ad effectiveness (e.g., image instead of video)

The market for content

- Monopolistic ISP connects CPs with consumers
- Bandwidth B
- Continuum of CPs indexed by $i \in [0, 1]$
  - Fraction μ have time sensitive content: delivered on time with probability $\gamma(\cdot)$
  - Fraction $1 - \mu$ have time insensitive content: no need for immediate delivery (always delivered on time)
- Homogeneous consumers derive utility u from each CP i whose content is delivered on time
- Each CP has one packet of content to deliver
- CPs charge prices p_i
- ISP charges subscription fee s and (if allowed) transmission fee t per packet (look at tiering later)

On-time delivery (1)

- Let A denote the total traffic on the (relevant part of the) network
- Probability that a given packet arrives on time:

$$\delta(A) = \min\left\{1, \frac{B}{A}\right\}$$

- Consider two systems of content delivery:
  - Single tier: best-effort principle
  - Two tiers: time-sensitive content prioritized
- Denote by α the average traffic generated by a time-sensitive CP
  - Single tier: $A = \mu\alpha + 1 - \mu$
  - Two tiers: on the priority lane, $A = \mu\alpha$

On-time delivery (2)

- Time-sensitive CPs can send their content more than once to increase the overall probability of delivery
- CP i chooses to send packet α_i times, with $\alpha_i \in \{0, 1, 2\}$
- Cost k per packet
- Overall probability of delivery of i's content:

$$\gamma(\alpha_i, A) = 1 - (1 - \delta(A))^{\alpha_i}$$

  $=$ probability that at least one packet sent by i arrives on time
- Assume $B < 1$. Thus, if $\alpha_i = 1$ for all i, not all traffic delivered on time in one-tiered system

Timing

1. ISP announces s and t.
2. Consumers choose whether to buy internet access from ISP.
3. Each CP chooses p_i and $\alpha_i \in \{0, 1, 2\}$.
4. Consumers choose to which CPs to connect.
5. Time insensitive content is delivered with probability 1. Time sensitive content is delivered with probability $\gamma(\alpha_i, A)$.
6. Consumers pay p_i to CPs whose content is delivered on time. CPs pay t for each unit of traffic carried by ISP.

We focus on symmetric subgame perfect equilibria.

Total surplus

- Total surplus from time-insensitive content is independent of delivery probabilities and given by $u - k$
- Total surplus from time-sensitive content can be written

$$W(\alpha) = \begin{cases} u \underbrace{\alpha}_{=\lambda_1} \gamma(1, A) - \alpha k & \text{for } \alpha \in [0, 1] \\ u\Big[\underbrace{(\alpha - 1)}_{=\lambda_2} \gamma(2, A) + \underbrace{(2 - \alpha)}_{=\lambda_1} \gamma(1, A)\Big] - \alpha k & \text{for } \alpha \in (1, 2]. \end{cases}$$

where

- λ_1: fraction of CPs sending 1 packet
- λ_2: fraction of CPs sending 2 packets

First best

- The first best always involves prioritization of time-sensitive content.

Lemma

The first-best traffic volume α^{FB} is such that there is no congestion and no duplication, i.e., each CP's content is sent at most once.

- A social planner prefers that all available content is delivered on time but some content unavailable rather than more content available but some of it delivered with delay
- Intuition: increasing α beyond $\hat{\alpha}_{dp}$ leaves amount of content delivered on time unchanged (but $\uparrow$ cost)

Second best

- Second best: time-insensitive content always sent, traffic volume of time-sensitive content (α) adjusted to maximize W

Lemma

The second-best traffic volume α^{SB} may involve congestion and duplication.

- With best-effort principle, planner does not eliminate congestion entirely
- Intuition: part of congestion cost borne by time-insensitive content

Equilibrium under net neutrality

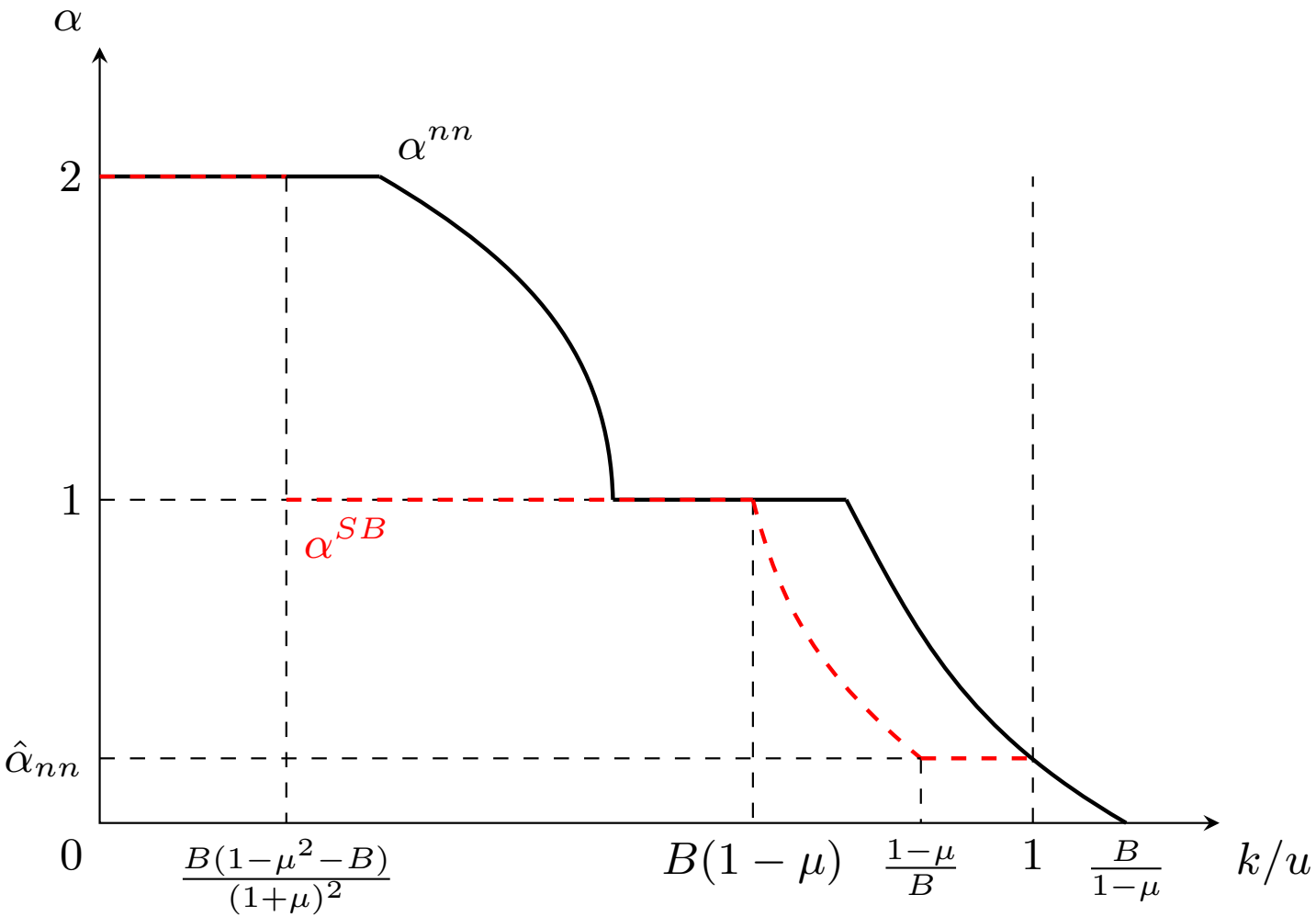
- Under net neutrality, $t = 0$
- Each CP sets $p_i = u$. Earns profit $\gamma(\alpha_i, A)u - k\alpha_i$.
- If they have internet access, consumers connect to all CPs
- They buy internet access if and only if $s \leq 0$. Hence, the ISP sets $s = 0$
- CP i compares profit from $\alpha_i \in \{0, 1, 2\}$ taking α as given:

α_i	profit
0	0
1	$u\gamma(1, A) - k$
2	$u\gamma(2, A) - 2k$

- Let α^{nn} denote the equilibrium traffic under net neutrality

Efficient vs equilibrium level of traffic

$B > (1-\mu^2)/(2-\mu)$ and $B > 1-\mu$



Traffic inflation

Proposition

The equilibrium level of traffic under net neutrality always exceeds the second-best level: $\alpha^{nn} \geq \alpha^{SB}$, with strict inequality for at least some range of k/u.

- CPs do not internalize the effect of increasing α_i on the aggregate level of traffic A
- This leads to inflation of traffic and generates excessive congestion

A uniform transmission fee

- Consider first the case where the ISP treats all traffic the same, but can charge $t > 0$ per unit of traffic
- CP i's problem becomes

$$\max_{\alpha_i} \gamma(\alpha_i, A)u - (k + t)\alpha_i$$

- The equilibrium gives us the inverse demand for traffic $t(\alpha)$ the ISP faces (assume in case of multiplicity the equilibrium most favorable to ISP is selected)
- The ISP solves

$$\max_{\alpha} \; t(\alpha)[\mu\alpha + 1 - \mu]$$

Profit-maximizing transmission fee

Proposition

The transmission fee that maximizes the ISP's profit is such that congestion is eliminated.

- At least a subset of time-sensitive CPs are shut down; active CPs send at most one packet
- Reminiscent of Anderson & de Palma (2009, RAND) information congestion paper: monopoly gatekeeper prices out congestion
- Note that in our context eliminating congestion entirely is generally not socially optimal

Deep packet inspection

- Deep packet inspection (DPI) allows ISP to prioritize time-sensitive content
- In times of shortage, all available bandwidth allocated to time-sensitive content
- Then, probability of on-time delivery higher for any given α
- Let α^{dp} denote the equilibrium level of traffic under deep packet inspection

DPI: a positive and a negative result

Proposition

If $B \geq \mu$, there exists an equilibrium under deep packet inspection in which the first-best level of traffic is transmitted irrespective of k/u, i.e., $\alpha^{dp} = \alpha^{FB} = 1$.

- If $B \geq \mu$, then all time-sensitive traffic arrives on time when each CP sends it once $\rightarrow$ no congestion, no incentive to deviate (there may be other equilibria)
- However, DPI often leads CPs to increase traffic compared to NN ($\rightarrow$ rent dissipation)

Proposition

The equilibrium probability of delivery for time-sensitive content can be lower under DPI than under NN.

Bandwidth tiering

- Two tiers of service: fast lane and slow lane
- ISP can allocate bandwidth B_f and B_s, respectively, with $B_f \geq B_s \geq 0$ and $B_f + B_s = B$
- Charges t_f and t_s, respectively
- It is optimal for the ISP to allocate all bandwidth to fast lane: $B_f = B$, $B_s = 0$
  - Time insensitive CPs don't care about speed
  - WTP of time sensitive CPs for fast lane decreases with B_s

Profit maximization with bandwidth tiering (1)

- The ISP's problem is

$$\max_{t_s,t_f} (1-\mu)t_s + \mu\alpha(t_f)t_f \quad \text{subject to} \quad t_s \leq t_f,$$

  where $\alpha(t_f)$ is demand for fast lane $=$ equilibrium traffic under DPI
- The constraint $t_s \leq t_f$ is binding at the optimum
- Thus, the ISP solves

$$\max_{\alpha} (1-\mu)t_f(\alpha) + \mu\alpha t_f(\alpha)$$

Profit maximization with bandwidth tiering (2)

Proposition

Under bandwidth tiering, the profit-maximizing transmission fee on the fast lane eliminates congestion.

- Qualitatively, same result as with uniform fee
- But here eliminating congestion is socially optimal
- Allowing the ISP to do bandwidth tiering and charge for the fast lane leads to the first-best outcome

Extensions

1. Congestion control techniques which reduce traffic:
   - Time-sensitive content also more bandwidth heavy: multiple packets to be delivered
   - Reducing the number of packets increases the probability of successful delivery:
     - Paying for compression
     - Reducing the quality of content by adjusting e.g. the video resolution
   - General insights robust to these alternative models:
     - Under net neutrality, socially insufficient adoption of traffic-reducing congestion control techniques (no internalization of effect of own decision on other CPs)
     - Bandwidth tiering with transmission fee on fast lane can increase efficiency
2. Heterogeneous user valuations of different content
   - Maintain symmetry
   - ISP can extract net consumer surplus and makes profit from connecting CPs with users

Conclusion

- Enforcing net neutrality may not be a good idea from point of view of network congestion: does not address externalities in traffic generation
- Departures from net neutrality can alleviate the problem:
  - Deep packet inspection may eliminate congestion, and thus incentive to inflate traffic
  - If ISP can charge transmission fees, will price out congestion
- But departures not unambiguously beneficial:
  - DPI can backfire if CPs respond by increasing traffic
  - Uniform transmission fees (without bandwidth tiering) will be set too high. Price cap can implement second-best traffic
- Bandwidth tiering can achieve first best