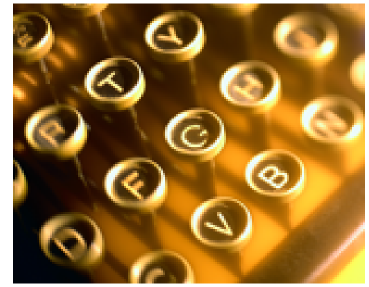
Game Theoretic Learning for Networked Control Systems

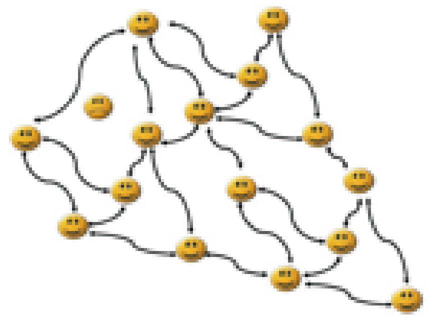
Jeff S Shamma
Georgia Institute of Technology

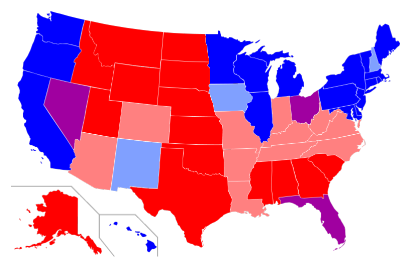
NecSys'09
24-26 September, 2009
Venice, Italy

Multagent scenarios

- Traffic
- Evolution of convention
- Social network formation
- Auctions & markets
- Voting
- etc
- Game elements (inherited):
  - Actors/players
  - Choices
  - Preferences

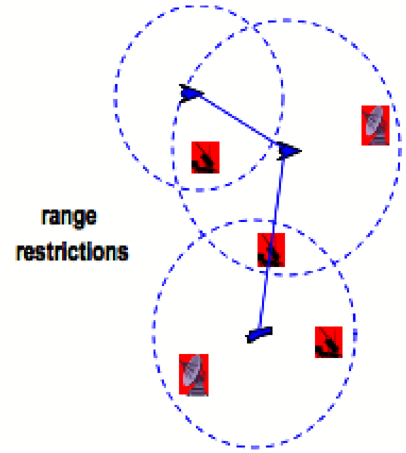


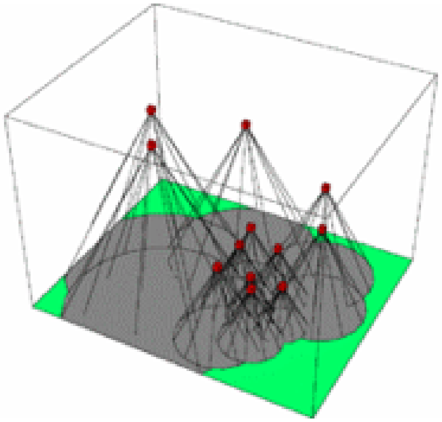


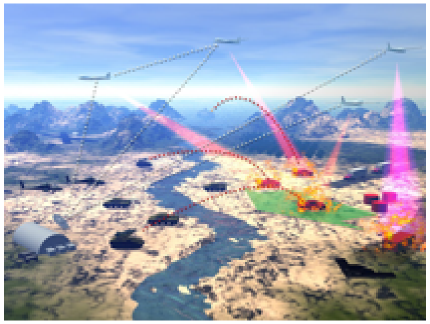


Descriptive Agenda

- Weapon-target assignment
- Data network routing
- Mobile sensor coverage
- Autonomous vehicle teams
- etc
- Game elements (designed):
  - Actors/players
  - Choices
  - Preferences







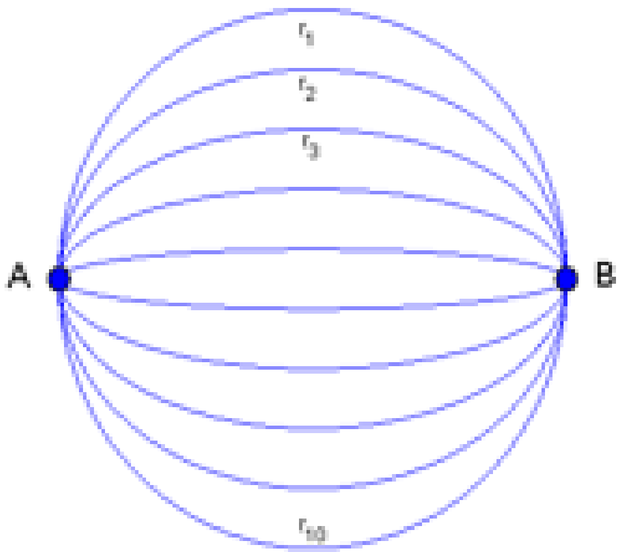
Prescriptive Agenda

Prescriptive agenda metrics

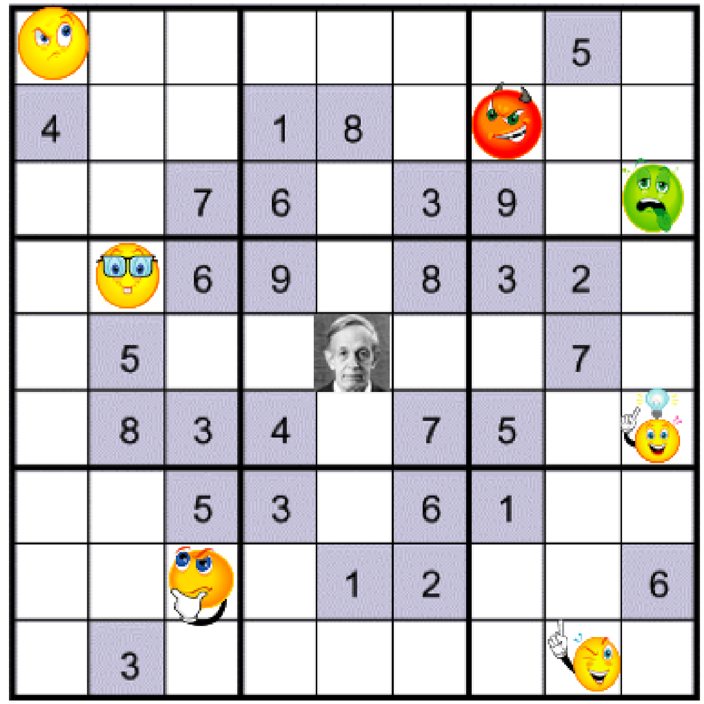
- Prescriptive agenda = distributed robust optimization
- *Choose* to address cooperation as noncooperative game
- Players are *programmable components* (vs humans)
- Must *specify*
  - Elements of game (players, actions, payoffs)
  - Learning algorithm
- Metrics:
  - Information available to agent?
  - Communications/stage?
  - Processing/stage?
  - Asymptotic behavior?
  - Global objective performance?
  - Convergence rates?

Outline

- Game theoretic learning
- Special class: Potential games
- Survey of algorithms
- Illustrations



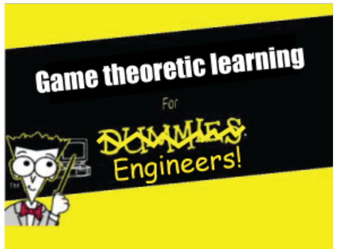
Distributed routing



Multi-agent sudoku

- Setup:
  - Players: $\{1, ..., p\}$
  - Actions: $a_i \in \mathcal{A}_i$
  - Action profiles:

$$(a_1, a_2, ..., a_p) \in \mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times ... \times \mathcal{A}_p$$



  - Payoffs: $u_i : (a_1, a_2, ..., a_p) = (a_i, a_{-i}) \mapsto \mathbf{R}$
  - Global objective: $G : \mathcal{A} \to \mathbf{R}$
- Action profile $a^* \in \mathcal{A}$ is a **Nash equilibrium** (NE) if for all players:

$$u_i(a_1^*, a_2^*, ..., a_p^*) = u_i(a_i^*, a_{-i}^*) \geq u_i(a_i', a_{-i}^*)$$

i.e., no *unilateral* incentive to change actions.

- Iterations:
  - $t = 0, 1, 2, ...$
  - $a_i(t) = \texttt{rand}(s_i(t)), \quad s_i(t) \in \Delta(\mathcal{A}_i)$
  - $s_i(t) = \mathcal{F}_i($available info at time $t)$
- Key questions: If NE is a descriptive outcome...
  - How could agents converge to NE?
  - Which NE?
  - Are NE efficient?

- Iterations:
  - $t = 0, 1, 2, ...$
  - $a_i(t) = \mathtt{rand}(s_i(t)), \quad s_i(t) \in \Delta(\mathcal{A}_i)$
  - $s_i(t) = \mathcal{F}_i($available info at time $t)$
- Key questions: If NE is a descriptive outcome...
  - How could agents converge to NE?
  - Which NE?
  - Are NE efficient?
- Focus shifted away from NE towards adaptation/learning

> *"The attainment of equilibrium requires a disequilibrium process"*
> K. Arrow
>
> *"Game theory lacks a general and convincing argument that a Nash outcome will occur."*
> Fudenberg & Tirole
>
> *"...human subjects are no great shakes at thinking either [vs insects]. When they find their way to an equilibrium of a game, they typically do so using trial-and-error methods."*
> K. Binmore

Survey: Hart, "Adaptive heuristics", 2005.

Game theoretic learning for prescriptive agenda?

- Approach: Use game theoretic learning to steer collection towards desirable configuration
- Informational hierarchy:
  - Action based: Players can observe the actions of others.
  - Oracle based: Players receive an aggregate report of the actions of others.
  - Payoff based: Players only measure online payoffs.
- Focus:
  - Asymptotic behavior
  - Processing per stage
  - Communications per stage

- For some $\phi : \mathcal{A} \to \mathbb{R}$

$$\phi(a_i, a_{-i}) - \phi(a_i', a_{-i}) > 0$$
$$\Leftrightarrow$$
$$u_i(a_i, a_{-i}) - u_i(a_i', a_{-i}) > 0$$

  i.e., potential function increases iff unilateral improvement.

- Features:
  - Typical of “coordination games”
  - Desirable convergence properties under various algorithms
  - Need not imply “cooperation” or $\phi = G$

- Distributed routing
  - Payoff = negative congestion. $c_r(\sigma_r)$
  - Potential function:

$$\phi = \sum_r \sum_{n=1}^{\sigma_r} c_r(n)$$

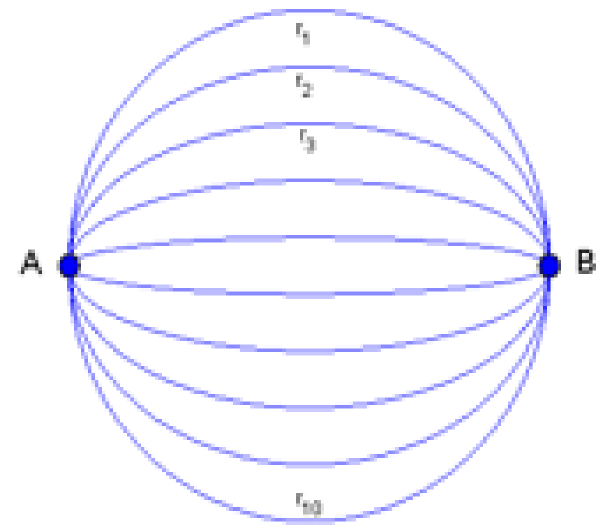
  - Overall congestion:

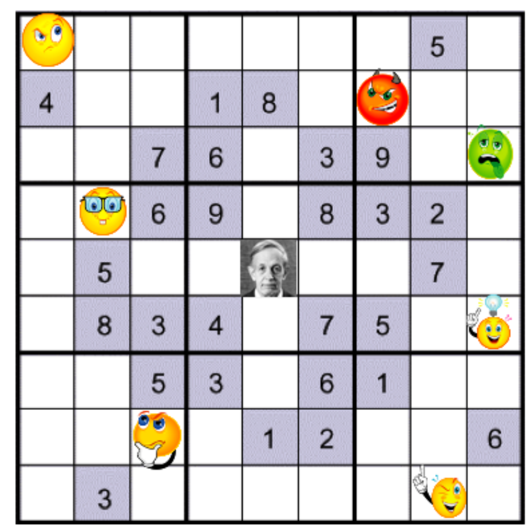
$$G = \sum_r \sigma_r c_r(\sigma_r)$$

  - **Note:** $\phi \neq G$
- Multiagent sudoku:

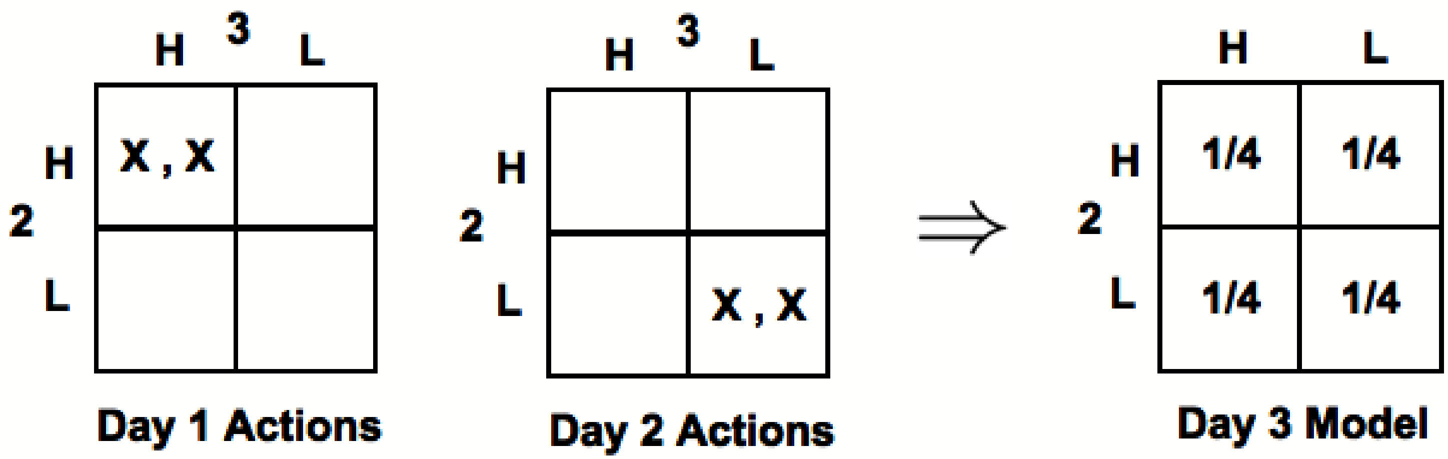
$u_i(a) =$#reps in row + #reps in column + #reps in sector

$$\phi(a) = \sum_i u_i(a)$$





- Each player:
  - Maintains empeirical frequencies (histograms) of other player actions
  - Forecasts (incorrectly) that others are playing randomly and independently according to empirical frequencies
  - Selects an action that maximizes expected payoff
- Bookkeeping is *action based*
- **Monderer & Shapley (1996)**: FP converges to NE in potential games.



- Viewpoint of driver 1 (3 drivers & 2 roads)
- Prohibitive-per-stage for large numbers of players with large action sets
  - Monitor all other players with IDs (cf., distributed routing)
  - Take expectation over large joint action space (cf., sudoku)

- Virtual payoff vector

$$U_i(t) = \begin{pmatrix} u_i(1, a_{-i}(t)) \\ u_i(2, a_{-i}(t)) \\ \vdots \\ u_i(m, a_{-i}(t)) \end{pmatrix}$$

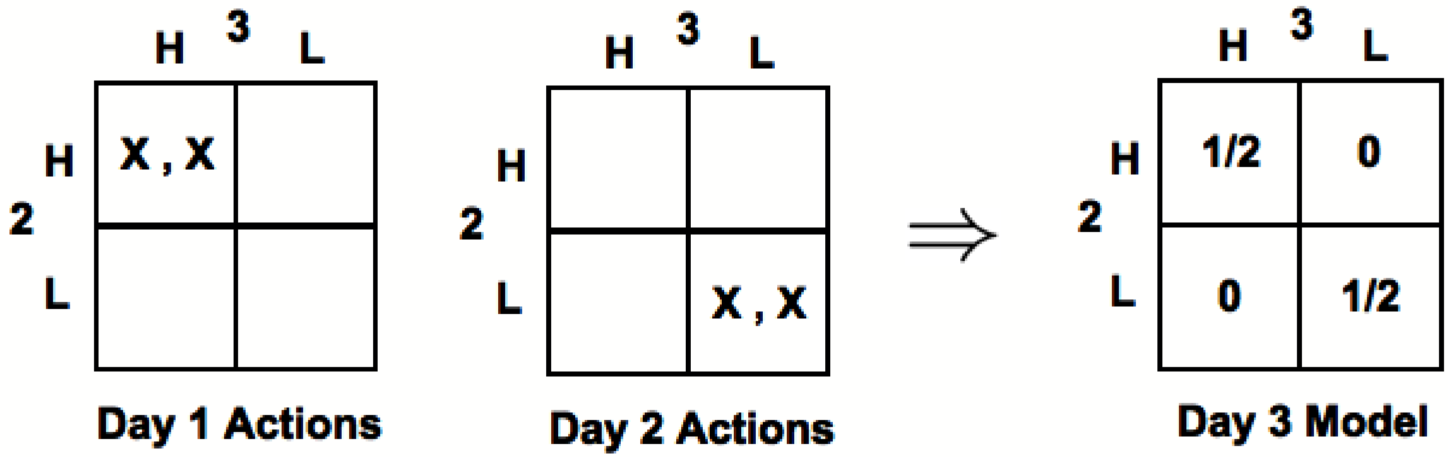
  i.e., the payofs that *could have* been obtained at time t

- Time averaged virtual payoff:

$$V_i(t+1) = (1-\rho)V_i(t) + \rho U_i(t)$$

- Stepsize ρ is either
  - Constant (fading memory)
  - Diminishing (true average), e.g., $\rho = \frac{1}{t+1}$
- Bookkeeping is *oracle based* (cf., traffic reports)

- JSFP algorithm: Each player
  - Maintains time averaged virtual payoff
  - Selects an action with maximal virtual payoff
  - OR repeats previous stage action with some probability (i.e., inertia)
- **Marden, Arslan, & JSS (2005)**: JSFP with inertia converges to a NE in potential games.



- Equivalent to best response to *joint actions* of other players
- Related to “no regret” algorithms
- Survey: Foster & Vohra, Regret in the online decision problem, 1999.

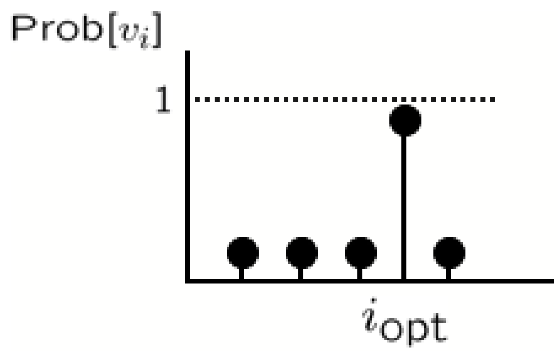
- Alternative algorithms offer more quantitative characterization of asymptotic behaviors.
- Preliminary: Gibbs distribution (cf., softmax, logit response)

$$\sigma(v;T) = \frac{1}{\mathbf{1}^{\mathrm{T}} e^{v/T}} e^{v/T} \in \Delta$$

e.g.,

$$\sigma(v_1, v_2; T) = \begin{pmatrix} \frac{e^{v_1/T}}{e^{v_1/T} + e^{v_2/T}} \\ \frac{e^{v_2/T}}{e^{v_1/T} + e^{v_2/T}} \end{pmatrix}$$

- As $T \downarrow 0$ assigns all probability to $\arg\max \{v_1, v_2, ..., v_n\}$



- At stage t
  - Player i is selected at random
  - Chosen player sets

$$a_i(t) = \text{rand}\left[\sigma\Big(u_i(1, a_{-i}(t-1)), ..., u_i(m, a_{-i}(t-1)); T\Big)\right]$$

  - Interpretation: Noisy best reply to previous joint actions
- Fact: SAP results in a Markov chain over joint action space $\mathcal{A}$ with a unique stationary distribution, μ.
- **Blume (1993)**: In (cardinal) potential games,

$$\mu(a) = \sigma(\phi(a); T) = \frac{e^{\phi(a)/T}}{\sum_{a' \in \mathcal{A}} e^{\phi(a')/T}}$$

- Implication: As $T \downarrow 0$, all probability assigned to potential maximizer.

- Motivation:
  - Reduced processing per stage
  - First step towards constrained actions
- At stage t:
  - Player i is selected at random
  - Chosen player compares $a_i(t-1)$ with randomly selected a_i'

$$a_i(t) = \text{rand}\left[\sigma(u_i(a_i(t-1), a_{-i}(t-1)), u_i(a_i', a_{-i}(t-1); T))\right]$$

- **Arslan, Marden, & JSS (2007)**: Binary SAP results in same stationary distribution as SAP.
- Consequence: Arbitrarily high steady state probability on potential function maximizer.

- Action evolution must satisfy: $a_i(t) \in \mathcal{C}(a_i(t-1))$
  - Limited mobility
  - Obstacles
- Algorithm: Same as before *except*

$$a_i' \in \mathcal{C}(a_i(t-1))$$

- **Marden & JSS (2008)**: Constrained SAP results in potential function maximizer being *stochastically stable*.
  - Arbitrarily high steady state probability on potential function maximizer
  - Does *not* characterize steady state distribution

Payoff based algorithms

- Action & oracle based algorithms require:
  - Explicit communications
  - Explicit representations of payoff functions
- Payoff based algorithms:
  - No (explicit) communication among agents
  - Only requires ability to *measure* payoff upon deployment

- Initialization of *baseline action* and *baseline utility*:

$$a_i^b(1) = a_i(0)$$
$$u_i^b(1) = u_i(a(0))$$

- Action selection:

$$a_i(t) = a_i^b(t) \text{with probability } (1-\epsilon)$$
$$a_i(t) \text{ is chosen randomly over } \mathcal{A}_i \text{ with probability } \epsilon$$

- Baseline action & utility update:

Successful Experimentation	*Unsuccessful Experimentation*	*No Experimentation*
$a_i(t) \neq a_i^b(t)$	$a_i(t) \neq a_i^b(t)$	$a_i(t) = a_i^b(t)$
$u_i(a(t)) > u_i^b(t)$	$u_i(a(t)) \leq u_i^b(t)$	
$\Downarrow$	$\Downarrow$	$\Downarrow$
$a_i^b(t+1) = a_i(t)$	$a_i^b(t+1) = a_i^b(t)$	$a_i^b(t+1) = a_i^b(t)$
$u_i^b(t+1) = u_i(a(t))$	$u_i^b(t+1) = u_i^b(t)$	$u_i^b(t+1) = u_i(a(t))$

- **Marden, Young, Arslan, & JSS (2007)**: For potential games,

$$\lim_{t \to \infty} \mathbf{Pr}\left[a(t) \text{ is a NE}\right] > p^*$$

  for any $p^* < 1$ with sufficiently small exploration rate ϵ.

- Suitably modified algorithm admits noisy utility measurements.

- How to assign individual payoff functions?
- Desirable features:
  - Induce “localization”
  - Have desirable NE
  - Produce potential game
- First attempt: Global utility
  - Set $u_i(a) = G(a)$ for all players.
  - Main disadvantage: Lack of localization
- Another issue: NE efficiency.
  - Global optimal:

$$a^* = \arg\max_{a \in \mathcal{A}} G(a)$$

  - Efficiency loss = “Price of Anarchy”

$$\text{PoA} = \min_{a \in \text{NE}} \frac{G(a)}{G(a^*)}$$

Illustration: Sensor placement

- Two sensors and two sectors:
  - Good sensor, g: $\mathbf{Pr}\left[\text{detect}\right] = 0.9$.
    Bad sensor, b: $\mathbf{Pr}\left[\text{detect}\right] = 0.1$.
  - High value sector: $H = 3$.
    Low value sector: $L = 2$.
- Optimal placement: $g = H$ & $b = L$.

g \ b	L	H
L	$0.91, 0.91$	$1.8, 0.3$
H	$2.7, 0.2$	$1.37, 1.37$

Equally shared: No pure NE

g \ b	L	H
L	$1.8, 0.2$	$1.8, 0.3$
H	$2.7, 0.2$	$2.7, 0.3$

Selfish: Optimal not NE

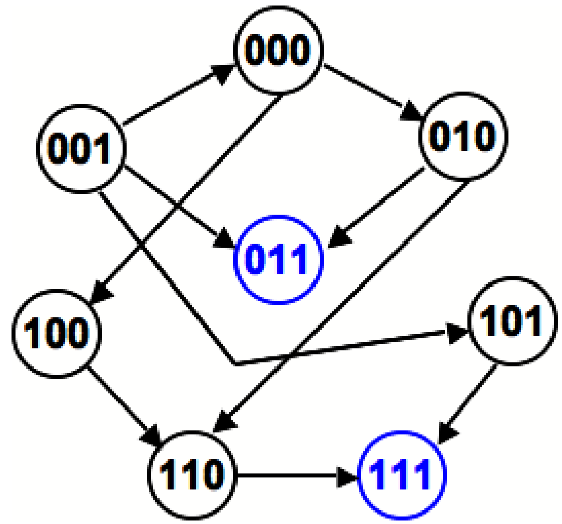
Marginal contribution payoff

- Introduce "null" action: $\emptyset$
- Interpretation: Context dependent
- Define:

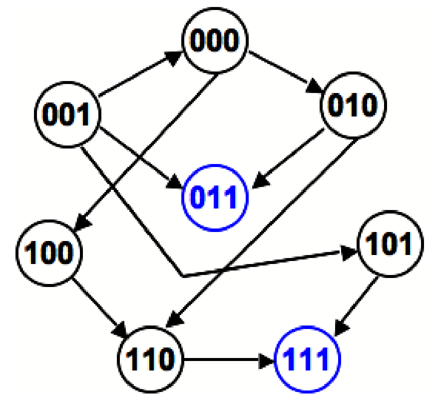
$$u_i(a_i, a_{-i}) = G(a_i, a_{-i}) - G(\emptyset, a_{-i})$$

  i.e., unilateral marginal contribution (also called "wonderful life utility" by Wolpert)
- Advantages:
  - Results in a potential game with $\phi = G$
  - Can induce "localization" effect in presence of spatial separatin
  - For sensor placement: Marginal contribution in selected cell

- Better reply graph
  - Nodes: joint actions
  - Directed edges: Better reply for *unilaterally* deviating player
- Illustration: 3 players, 2 moves each
- Features:
  - Potential function increases along edges
  - NE iff no outgoing edges



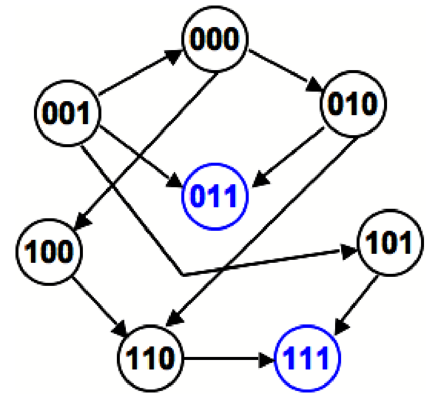
- Recall max regret with inertia:
  - Players monitor regret vector & choose maximal regret action
  - OR repeat previous action with some probability
  - Regret maximizer is not best reply to previous stage
- A path to NE that occurs with $\delta > 0$ probability:
  - Players linger (inertia)
  - Eventually, regret maximizer = best reply to joint action
  - Single player deviates if not NE
  - Repeat
- NE + lingering implies permanent NE
- Cannot avoid NE path indefinitely



- Recall (binary) SAP
  - Single agent, randomly selected
  - Uses Gibbs distribution to select next action
- Features:
  - Node hops not limited to better replies (softmax)
  - Better replies have higher probabilities
- Detailed balance equation:

$$\mathbf{Pr}\left[a \to a'\right] \mu(a) = \mathbf{Pr}\left[a' \to a\right] \mu(a')$$

  (stronger condition than stationary distribution)
- Proof: Transition probabilities under SAP satisfy detailed balance equation with Gibbs distribution for *potential games*.



- Recall *stochastic stability* definition:
  - Let P^ϵ denote the transition probability matrix of an irreducible & aperiodic Markov chain.
  - Let μ^ϵ be the (unique) stationary distribution for P^ϵ
  - A state, x, is **stochastically stable** if

$$\liminf_{\epsilon \to 0} \mu^\epsilon(x) > 0$$

- Implication: Increasing probability of being in stochastically stable state with decreasing ε.
- Utilization:
  - Payoff based experimentation: NE are only stochastically stable baseline actions.
  - Constrained SAP: Potential maximizers are only stochastically stable joint actions.

- **Young (1993)**: To determine stochastic stability
  - View learning dynamics as ϵ perturbation of reference ($\epsilon = 0$) Markov chain
  - Divide reference Markov chain into recurrence classes (typically Nash equilibria)
  - Define *resistance* to transition between recurrence classes:

$$0 < \lim_{\epsilon \downarrow 0} \frac{P^{\epsilon}_{ij}}{\epsilon^{r(i \to j)}} < \infty$$

  - Form *stochastic potential* for each recurrence class
  - Minimal stochastic potential implies stochastic stability
- Trivial illustration:
  - Perturbed & Reference Markov Chain:

$$P^{\epsilon} = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon^2 & 1-\epsilon^2 \end{pmatrix} \quad P^0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

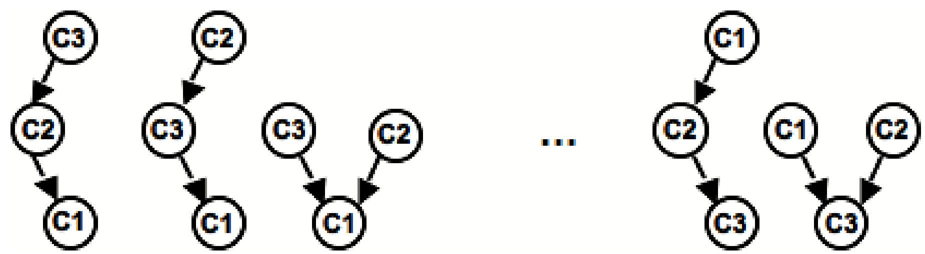
  - Resistances:

$$0 < \lim_{\epsilon \downarrow 0} \frac{P^{\epsilon}_{ij}}{\epsilon^{r(i \to j)}} < \infty$$

$$r(1 \to 2) = 1 \quad \& \quad r(2 \to 1) = 2$$

  - Stochastically stable state: 2

- Analytical utilization:
  - Do *not* build all trees



  - Show that one tree has lower stochastic potential than another

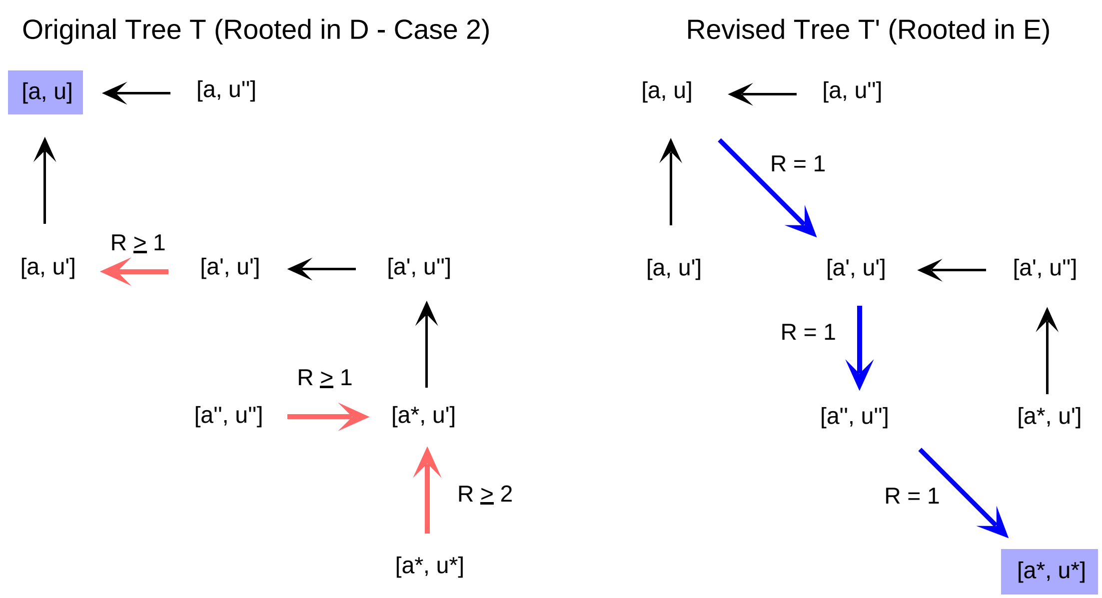
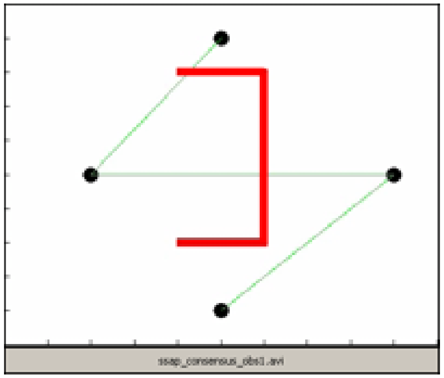


Illustration: Rendezvous with obstacles

- Assume undirected connected constant graph (can be generalized)
- Global objective:

$$G(a_i, a_{-i}) = -\frac{1}{2}\sum_k \sum_{j\in\mathcal{N}_k} |a_k - a_j|$$



- Global objective without agent i

$$G(\emptyset, a_{-i}) = -\frac{1}{2}\sum_{k\neq i} \sum_{j\in\mathcal{N}_k\setminus i} |a_k - a_j|$$

- Marginal contribution utility:

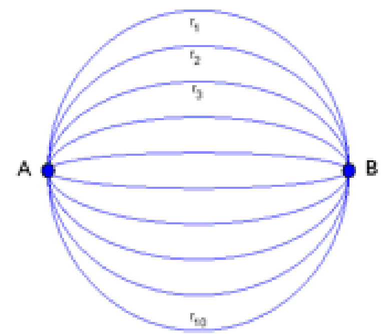
$$u_i(a_i, a_{-i}) = G(a_i, a_{-i}) - G(\emptyset, a_{-i}) \qquad = -\sum_{j\in\mathcal{N}_i} |a_i - a_j|$$

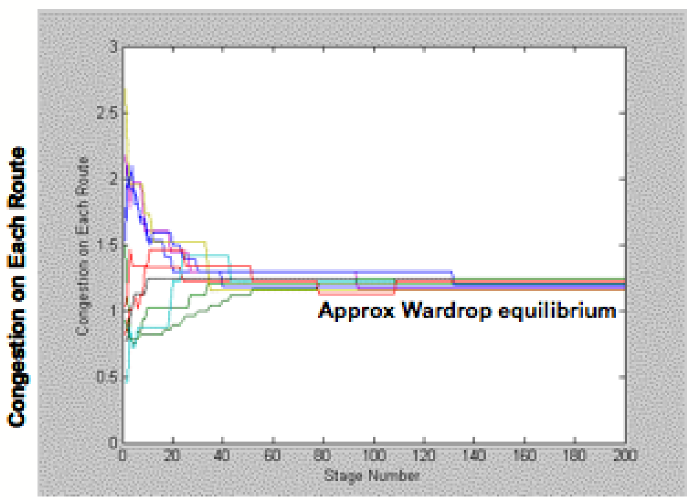
- Apply constrained SAP...

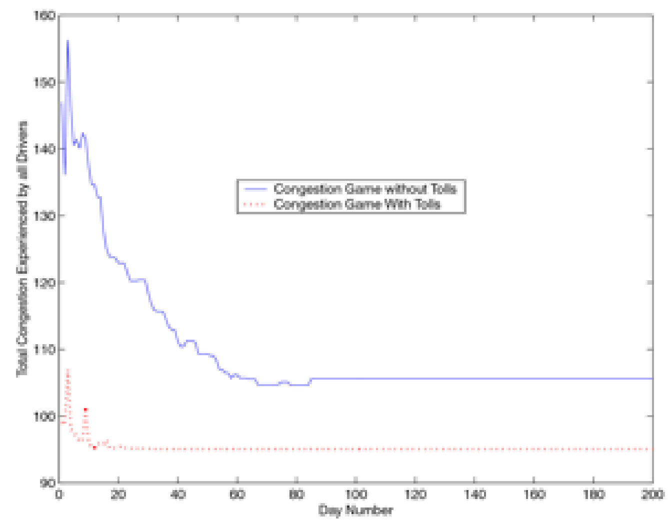
- Setup: 10 parallel roads. 100 vehicles.
- Marginal contribution utility using overall congestion induces "tolls"

$$\tau_r(k) = (k-1) \cdot (c_r(k) - c_r(k-1))$$

- Apply max regret with intertia...







- Recap:
  - Descriptive vs prescriptive
  - Action/Oracle/Payoff based algorithms
  - NE or potential function maximization
  - Potential games & payoff design
- Future work:
  - Convergence rates
  - Exploiting prescriptive setting
  - Agent dynamics
  - Control theory and *descriptive* agenda