FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

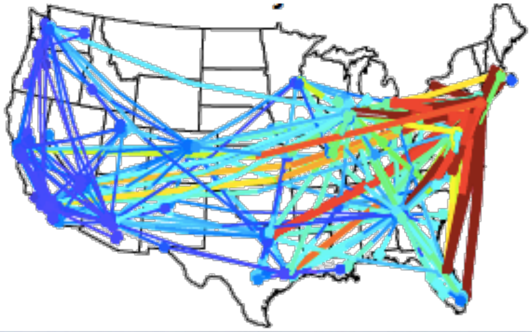
Data-Driven Modeling and Optimization Algorithms for h-CPS

Hamsa Balakrishnan
(with Jacob Avery, Michael Kasperski & Varun Ramanujam)
Massachusetts Institute of Technology

Berkeley UNIVERSITY OF CALIFORNIA
Massachusetts Institute of Technology
NSF
MICHIGAN
VANDERBILT UNIVERSITY

Most critical infrastructures are *h-CPS*

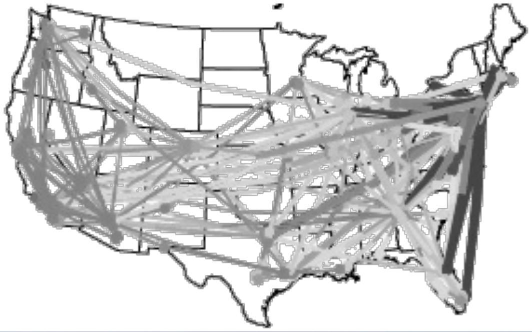
* Cyber + Physical + ***human*** (decision makers)
  * How do we model decision processes?
  * How do we estimate utility functions?
  * How do we predict system performance, when it depends on human decision-makers?
* Large-scale networks
  * Disruptions propagate through the system
* Multi-stakeholder systems
  * Optimization algorithms for resource allocation
  * Incentives for information-sharing



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Most critical infrastructures are *h-CPS*

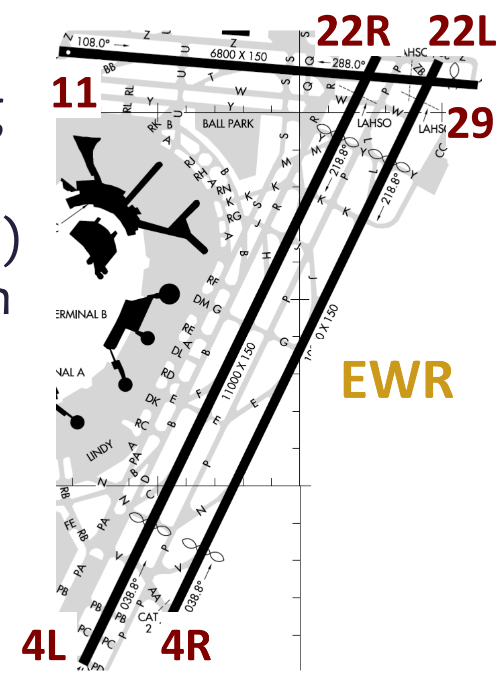
- Cyber + Physical + ***human*** (decision makers)
  - How do we model decision processes?
  - How do we estimate utility functions?
  - How do we predict system performance, when it depends on human decision-makers?
- Large-scale networks
  - Disruptions propagate through the system
- Multi-stakeholder systems
  - Optimization algorithms for resource allocation
  - Incentives for information-sharing



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Modeling human decision processes

* Key challenge in h-CPS is modeling/predicting the behavior of the human participants
  * Discrete-choice models
    * Assuming decision-makers are rational, estimate their utility functions
    * Estimate relative weightings of different influencing factors
    * Use operational data (i.e., observations of decisions) to determine maximum-likelihood model of decision process
  * Approach demonstrated on airport configuration selection
    * Which runways are used for which operations
    * Primary driver of airport capacity



Discrete-choice models: Utility function

* Rich literature in transportation demand analysis (Ben-Akiva & Lerman 1985)
* Decision-maker chooses utility-maximizing option (from a set)
* Utility function is modeled as a linear function of the independent variables plus an error term

$$U_i = \underbrace{(\alpha_i + \beta_i \cdot X_i)}_{\text{Observed component, } V_i} + \underbrace{\epsilon_i}_{\text{Unobserved error}}$$

* For each observation, assume that the decision-maker chooses the alternative that maximizes utility, i.e., the choice c_j such that

$$j = \arg\max_{i:c_i \in C} U_i$$

* Different models arise from assuming different forms of error term
  * Most widely used class of models assumes that the errors are independent and identically Gumbel distributed
  * Logistic Probability Unit, or Logit models

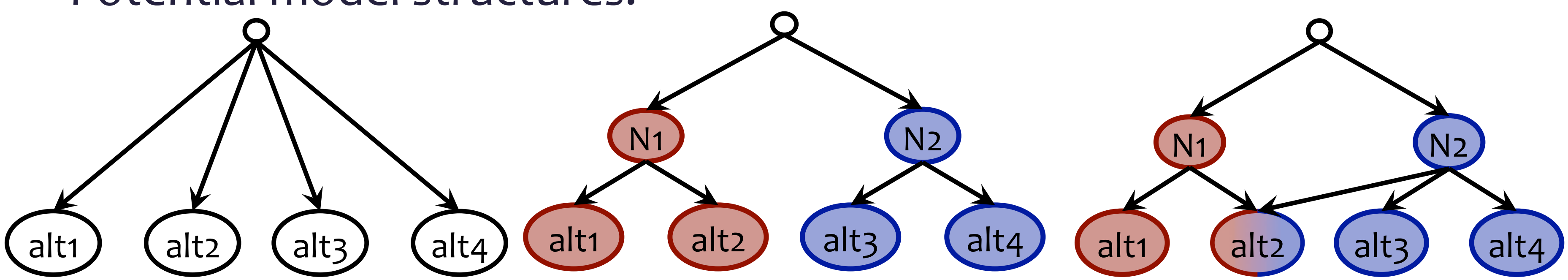
FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Discrete-choice models: Structure

* Alternatives (error terms) may not necessarily be independent

$$U_i = \underbrace{(\alpha_i + \beta_i \cdot X_i)}_{\text{Observed component, } V_i} + \underbrace{\epsilon_i}_{\text{Unobserved error}}$$

* Potential model structures:



Multinomial Logit (MNL) Nested Logit (NL) Cross Nested Logit (CNL)

* Alternatives within the same nest have correlated error terms
  * NL example: $V_{\mathrm{N1}} = \frac{1}{\mu_{\mathrm{N1}}} \log \sum_{j: c_j \in \{\mathrm{alt1}, \mathrm{alt2}\}} e^{\mu_{\mathrm{N1}} V_j}$; $P(\mathrm{N1} | \{\mathrm{N1}, \mathrm{N2}\}) = \frac{e^{V_{\mathrm{N1}}}}{e^{V_{\mathrm{N1}}} + e^{V_{\mathrm{N2}}}}$

$$P(\mathrm{alt1} | \mathrm{N1}) = \frac{e^{\mu_{\mathrm{N1}} V_{\mathrm{alt1}}}}{\sum_{j: c_j \in \{\mathrm{alt1}, \mathrm{alt2}\}} e^{\mu_{\mathrm{N1}} V_j}}$$

Maximum-likelihood estimation

* Explanatory variables (X_i) of the utility function are determined iteratively
* For a given functional form of the utility function, the likelihood function of a given set of observations (over N time periods, say) is

$$\mathscr{L}(\alpha, \beta) = P((c_1|C_1) \bigcap \ldots \bigcap (c_N|C_N)|\alpha, \beta, X)$$

where c_n is the choice observed at time n, and C_n is the set of options
* Assuming that the observed choices at each time are conditionally independent given the explanatory variables, we get

$$\mathscr{L}(\alpha, \beta) = \prod_{i=1}^{N} P(c_i|C_i)$$

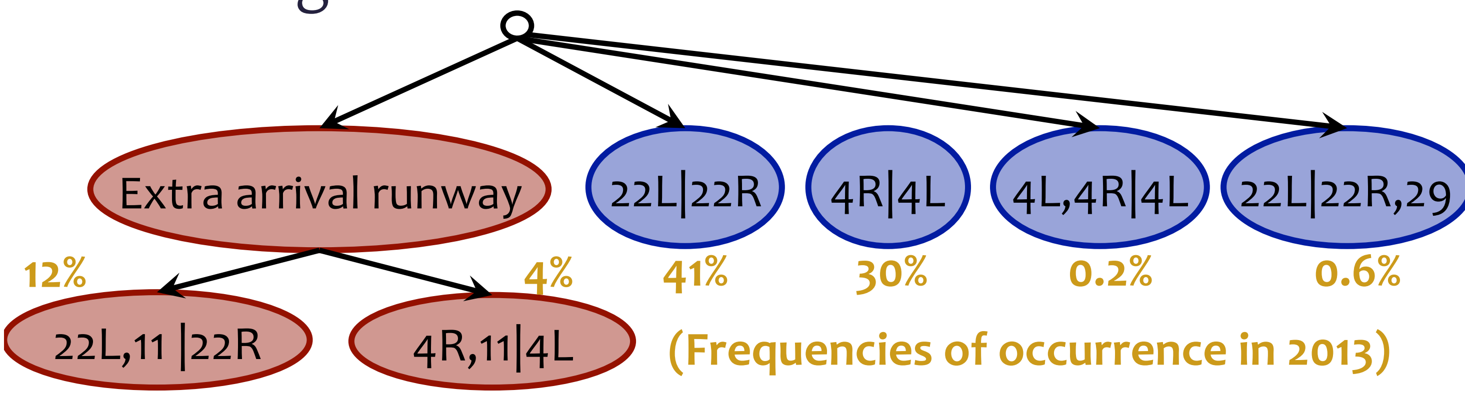
$$(\widehat{\alpha}, \widehat{\beta}) = \arg\max_{\alpha,\beta} \mathscr{L}(\alpha, \beta)$$

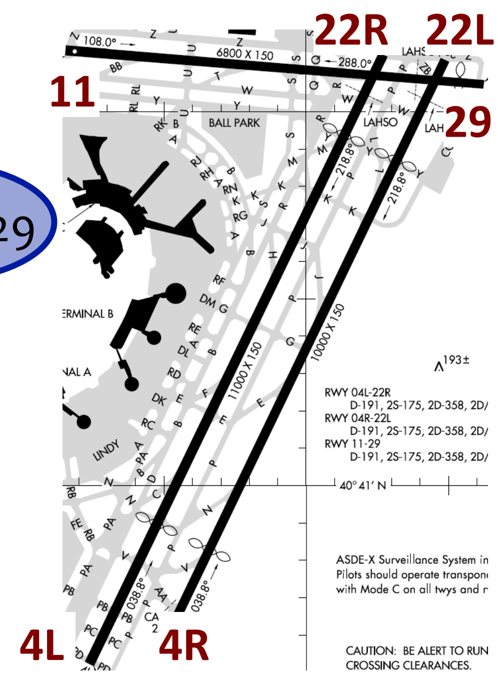
* Nonlinear optimization problem (Bierlaire, 2003)
* Structure (MNL/NL/CNL) determined by checking statistical significance (Hausman-McFadden '84)

FORCES FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Results: Newark (EWR) airport case study

* Training data: 2011

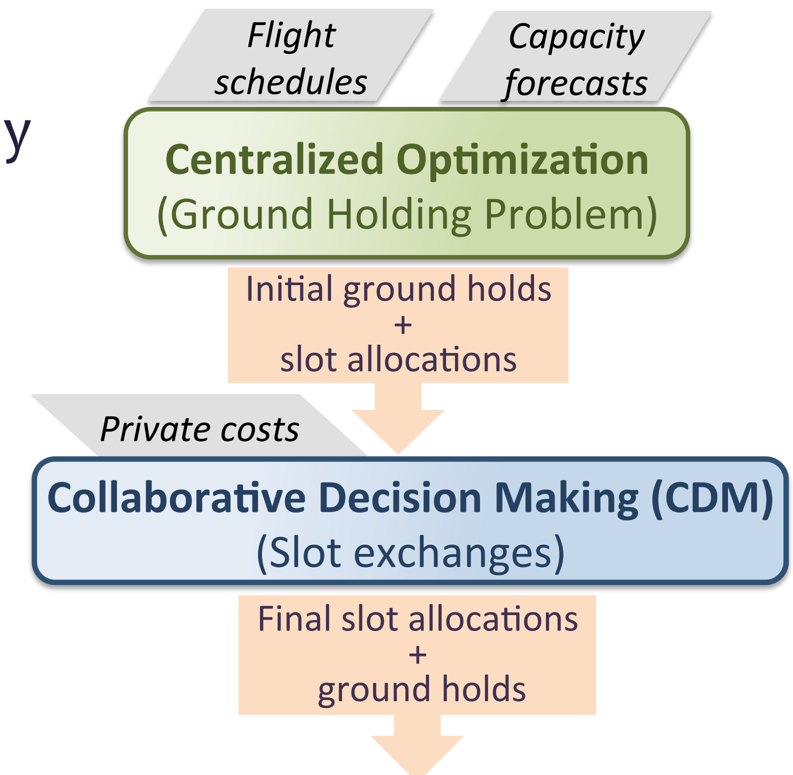




  * Only consider configurations used >50 times/year
* Validation data: 2013
  * Prediction accuracy: 73% (max possible: 89%)
  * 100% accuracy in distinguishing between use of the 22's and the 4's
* Currently extending approach to LGA and SFO
* Other applications

Resource Allocation: Optimization + Collaborative Decision Making

- Centralized optimization generally **assumes homogeneous delay costs**
- Airport capacity is uncertain, especially a few hours ahead of time
- Stochastic optimization formulations:
  - Static: Single-stage stochastic Integer Program (IP)
  - Dynamic: Multi-stage stochastic IP, differentiates between flights of different durations
  - *Hybrid:* Multi-stage stochastic IP, but does not differentiate between flights of different durations



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Static Ground Holding Problem

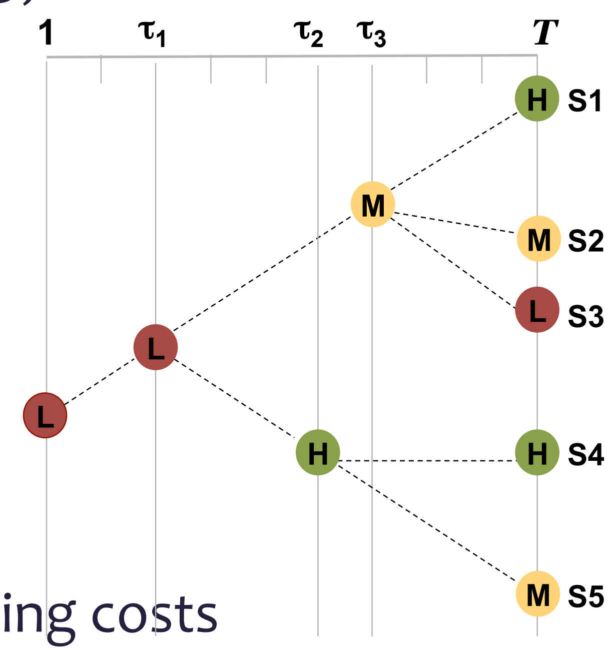
* Single-stage stochastic IP (Richetta & Odoni 1993)

$$\text{Minimize} \sum_{n=0}^{K} C_{g,n}\Big(\sum_{t=1}^{T-n} A^{\text{gq}}_{t,t+n}\Big) + \sum_{q\in Q} \pi_q\Big(C_a \sum_{t=1}^{T} A^{\text{aq}}_{q,t}\Big)$$

$$\text{subject to} \sum_{j=t}^{t+K} A^{\text{gq}}_{t,j} = A^{\text{dem}}_{t},\ \forall t \in \{1,..,T\}$$

$$A^{\text{aq}}_{q,t} \geq \sum_{j=t-K}^{t} A^{\text{gq}}_{j,t} + A^{\text{aq}}_{q,t-1} - A^{\text{cap}}_{q,t},\ \forall t \in \{1,..,T\}, q \in Q$$

$$A^{\text{gq}}_{t,j}, A^{\text{aq}}_{q,t} \in \mathbb{Z}^{+},\ \forall t,j \in \{1,..,T\}, q \in Q$$



* LP relaxation is integer-optimal if ground-holding costs are marginally non-decreasing (Kotnyek & Richetta 2006)

Dynamic Ground Holding Problem

* Multi-stage stochastic IP (Mukherjee & Hansen 2007)

$$\text{Minimize} \sum_{q\in Q}\pi_q\Big[\sum_{f\in F}\Big(\sum_{t=\text{arr}_f}^{\text{arr}_f+K} C_{g,t-\text{arr}_f} X^q_{f,t}\Big) + \Big(C_a \sum_{t=1}^{T} A^{\text{aq}}_{q,t}\Big)\Big]$$

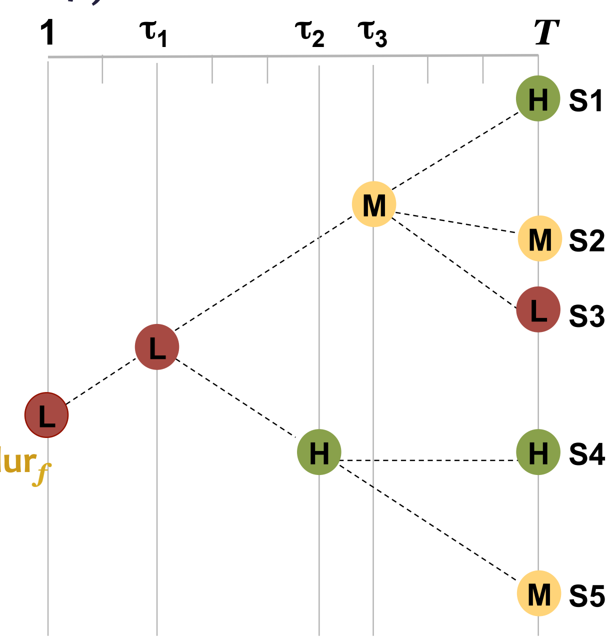
$$\text{subject to} \sum_{t=\text{arr}_f}^{\text{arr}_f+K} X^q_{f,t} = 1,\ \forall q \in Q, \forall f \in F$$

$$A^{\text{aq}}_{q,t} \geq \sum_{f\in F} X^q_{f,t} + A^{\text{aq}}_{q,t-1} - A^{\text{cap}}_{q,t},\ \forall t \in \{1,..,T\}, q \in Q$$

$$X^{q_1}_{f,t} = X^{q_2}_{f,t},\ \forall q_1, q_2 \in G_{t-\text{dur}_f}$$

← set of feasible scenarios at time t-dur$_f$

$$X^q_{f,t} \in \{0,1\},\ A^{\text{aq}}_{q,t} \in \mathbb{Z}^+,\ \forall t \in \{1,..,T\},\ \forall q \in Q,\ \forall f \in F$$



* In general, LP relaxation solution is not integer-optimal
* $\mathcal{O}(FT^2+T^2)$ integer decision variables

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Hybrid Ground Holding Problem

- Multi-stage stochastic IP (Ramanujam & Balakrishnan CDC 2014, submitted)

$$\text{Minimize} \sum_{q\in Q} \pi_q \left(\sum_{n=0}^{K} C_{g,n} \sum_{t=1}^{T-n} X^q_{t,t+n} + C_a \sum_{t=1}^{T} A^{\text{aq}}_{q,t} \right)$$

$$\text{subject to} \sum_{j=t}^{t+K} X^q_{t,j} = A^{\text{dem}}_t, \forall t \in \{1,..,T\}, q \in Q$$

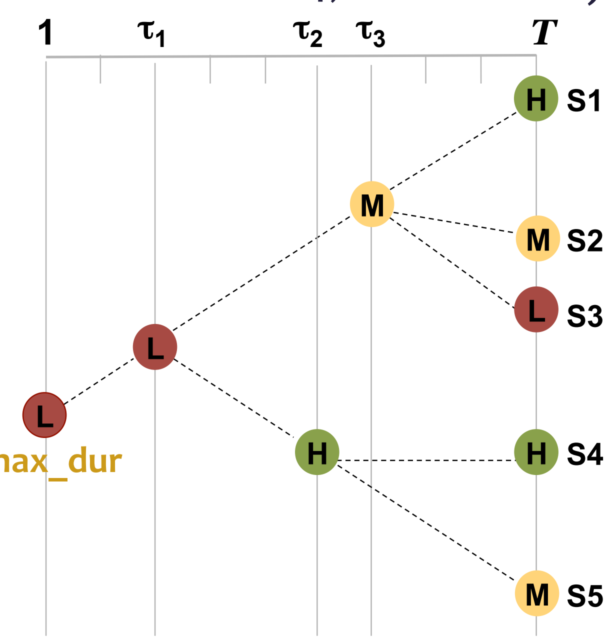
$$A^{\text{aq}}_{q,t} \geq \sum_{j=t-K}^{t} X^q_{j,t} + A^{\text{aq}}_{q,t-1} - A^{\text{cap}}_{q,t}, \forall t \in \{1,..,T\}, q \in Q$$

$$X^{q1}_{t,j} = X^{q2}_{t,j}, \forall q_1, q_2 \in G_{t-\text{max\_dur}}$$

set of feasible scenarios at time *t*-max_dur

$$X^q_{t,j} \in \mathbb{Z}^+, \forall t,j \in \{1,..,T\}, q \in Q$$

$$A^{\text{aq}}_{q,t} \in \mathbb{Z}^+, \forall t \in \{1,..,T\}, \forall q \in Q.$$



- In general, $\mathcal{O}(T^3)$ integer decision variables

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Properties of the Hybrid Ground Holding Formulation

* Marginally non-decreasing ground-holding costs

LEMMA 1. *The hybrid stochastic SAGHP formulation yields an optimal solution with integer values for all variables* $X^q_{a,b}$ $(\forall q \in Q; a, b \in \{1, .., T\})$ *if the queue length variables* $(A^{aq}_{q,t}\ \forall q \in Q, t \in \{1, .., T\})$ *are constrained to have integer values, and the ground-holding costs are marginally non-decreasing (i.e.,* $C_{g,n+1} - C_{g,n} \geq C_{g,n} - C_{g,n-1}\ \forall n$*).*

* i.e., $\mathcal{O}(T^2)$ **integer variables instead of** $\mathcal{O}(T^3)$

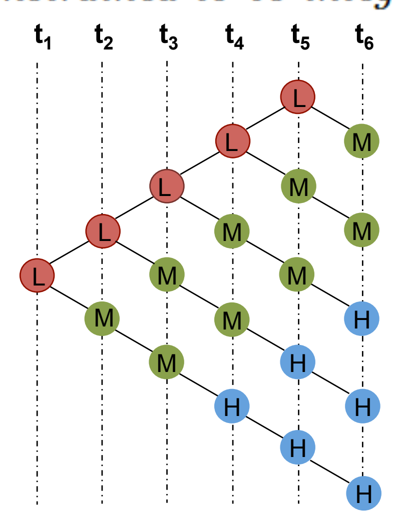
[Ramanujam & Balakrishnan CDC 2014, submitted]

Properties of the Hybrid Ground Holding Formulation

* Marginally non-decreasing ground-holding costs + special scenario tree structure

LEMMA 2. *Given marginally non-decreasing ground-holding cost coefficients* $C_{g,n+1} - C_{g,n} \geq C_{g,n} - C_{g,n-1}$, $\forall n$, *and a capacity scenario tree forecast with sequentially non-decreasing capacity scenarios and sole element of uncertainty being time of improvement from lowest capacity state, the hybrid stochastic SAGHP formulation is guaranteed to have an integral optimum solution if the queue length variables for scenario* T *(i.e.,* $A^{\text{aq}}_{T,t}\ \forall t \in \{1,..,T\}$*) are constrained to be integers.*

* i.e., **$\mathcal{O}(T)$ integer variables instead of $\mathcal{O}(T^3)$**



[Ramanujam & Balakrishnan CDC 2014, submitted]

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Optimization formulations of Collaborative Decision Making

Intra-airline substitution

$$\text{Minimize} \quad \sum_{f_1 \in F_a} \sum_{f_2 \in F_a} C_{f_1,f_2} X_{f_1,f_2}$$

$$\text{subject to:} \quad \sum_{f_1 \in F_a} X_{f_1,f_2} = 1, \; \forall f_2 \in F_a$$

$$\sum_{f_2 \in F_a} X_{f_1,f_2} = 1, \; \forall f_1 \in F_a$$

$$X_{f_1,f_2} \leq \text{feas}_{f_1,f_2}, \; \forall f_1, f_2 \in F_a$$

$$X_{f_1,f_2} \in 0, 1 \, \forall f_1, f_2 \in F_a$$

Inter-airline substitution

$$\text{maximize} \quad \sum_{q \in Q} \pi_q \left[\sum_{f \in F \setminus c} \sum_{r=1}^{T} \mathcal{B}(r) Y^q_{f,r} - M d^q_c \right]$$

subject to:

$$\sum_{t=\text{ETA}_f}^{\text{ETA}_f + K} X^q_{f,t} = 1, \; \forall q \in Q, \forall f \in F$$

$$A^{\text{aq}}_{q,t} \geq \sum_{f \in F} X^q_{f,t} + A^{\text{aq}}_{q,t-1} - A^{\text{cap}}_{q,t}, \; \forall t \in \{1, .., T\}, q \in Q$$

$$X^{q_1}_{f,t} = X^{q_2}_{f,t}, \; \forall q_1, q_2 \in G_{\text{ETA}_f - \text{max\_dur}},$$

$$d^q_c = \sum_{t=1}^{T} t X^q_{c,t} - k, \; \forall q \in Q,$$

$$\sum_{t=1}^{T} t X^q_{f,t} \leq \text{arr}^q_f, \; \forall q \in Q, f \in F \setminus c$$

$$A^{\text{aq}}_{q,t} \leq Aq^{\text{aq,orig}}_{q,t}, \; \forall q \in Q, t \in \{1, .., T\}$$

$$Y^q_{f,r} = X^q_{f, \text{arr}^q_f - r + 1}, \; \forall q \in Q, f \in F \setminus c, r \in \{1, .., T\}$$

$$X^q_{f,t} \in \{0, 1\}, \; d^q_c \geq 0, \; \forall t \in \{1, .., T\}, \forall q \in Q, \forall f \in F$$

Comparison of Ground Holding Problem Formulations

	Static	Hybrid	Dynamic
Pre-CDM delay cost	High (Worst)	Medium	Low (Best)
Benefit from CDM	High (Best)	Medium	Low (Worst)
Equity	High (Best)	Medium	Low (Worst)
Tractability	High (Best)	Medium	Low (Worst)
Ease of implementation	High (Best)	Medium	Low (Worst)

[Ramanujam & Balakrishnan CDC 2014, submitted]

Summary

* Modeling human-driven decision processes is difficult, since utility functions are often not formally codified
  * Discrete-choice models present a way to determine utility functions as well as model structure
  * Data-driven models (descriptive, rather than prescriptive)
  * Maximum-likelihood estimation
* Multi-stakeholder optimization is a critical challenge in h-CPS
  * Different optimization formulations present tradeoffs in terms of
    * Computational tractability
    * System-optimal benefits
    * Incentives for participation
    * Incentives for information-sharing