

Algebraic Property Graphs

Joshua Shinavier¹ Ryan Wisnesky² Joshua G. Meyers²

¹Uber Technologies, Inc.¹

²Conexus AI

July 19, 2022

¹For the toolkit, applications, and original formulation of APG. The author's current affiliation is LinkedIn Corporation.

Abstract: We present a case study in applied category theory written from the point of view of an applied domain: the formalization of the widely-used *property graphs* data model in an enterprise setting using elementary constructions from type theory and category theory, including limit and co-limit sketches. Observing that *algebraic data types* are a common foundation of most of the enterprise schema languages we deal with in practice, for graph data or otherwise, we introduce a type theory for *algebraic* property graphs wherein the types denote both algebraic data types in the sense of functional programming and join-union E/R diagrams in the sense of database theory. We also provide theoretical foundations for graph transformation along schema mappings with by-construction guarantees of semantic consistency. Our data model originated as a formalization of a data integration toolkit developed at Uber which carries data and schemas along composable mappings between data interchange languages such as Apache Avro, Apache Thrift, and Protocol Buffers, and graph languages including RDF with OWL or SHACL-based schemas.

Design considerations:

- ▶ Formalize the labeled property graph data model
- ▶ Formalize data interchange languages including Protobuf, Thrift, and Avro
- ▶ Serve as a standard, unifying data model for Uber
- ▶ Serve as a bridge between graph and non-graph data sources

Applications:

- ▶ Basis for reusable, standardized data types
- ▶ Mapping of common types into multiple data languages
- ▶ Mapping of schemas and data across languages
- ▶ Language-agnostic schema and data validation
- ▶ Graph (PG and RDF) construction

Graph data models

- ▶ Property Graphs: edge-labeled and vertex-labeled graphs with properties. See Apache TinkerPop and graph databases such as Neo4j
- ▶ RDF: W3C standard graph data model. Graphs are sets of subject/predicate/object triples. Various schema formalisms (RDFS, OWL, SHACL, etc.)
- ▶ Hypergraphs: variations on the mathematical concept of a hypergraph. Less common than property graphs, but more compatible with relational databases

Property graph elements

- ▶ Vertex, edge: as in a classic labeled, directed multigraph
- ▶ Vertex property, edge property: key/value pair attached to a vertex or edge
- ▶ Metaproperty: property attached to another property
- ▶ etc.

Algebraic data types in practice

| | Apache Thrift | Apache Avro | Protocol Buffers v3 | GraphQL SDL |
|---------------|---------------|-------------|---------------------|-------------|
| product types | yes | yes | yes | yes |
| sum types | yes | yes | yes | yes |
| interfaces | | | | yes |
| enumerations | yes | yes | yes | yes |
| optionals | yes | yes | | yes |
| typedefs | yes | | | |
| defaults | yes | yes | yes | yes |
| constants | yes | | | |
| lists/arrays | yes | yes | yes | yes |
| maps | yes | yes | yes | |
| sets | yes | | | |

Figure: Comparison of selected data exchange languages

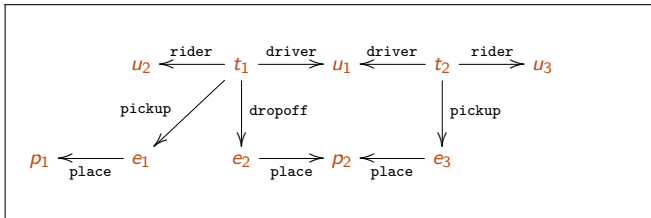
APGs: Definition 1

Fix a set P of primitive types p with extensions $PV(p)$.

An APG schema S consists of:

- ▶ $\mathcal{L} : \mathbf{Set}$
- ▶ $\sigma : \mathcal{L} \rightarrow \mathcal{T}$, where

$$\mathcal{T} \ni t ::= 0 \mid 1 \mid t_1 + t_2 \mid t_1 \times t_2 \mid \text{Prim } p \ (p \in P) \\ \mid \text{Lbl } l \ (l \in \mathcal{L})$$



Let $P := \{\text{Integer}\}$, $PV(\text{Integer}) = \mathbb{Z}$.

Example schema $S = (\mathcal{L}, \sigma)$:

$$\begin{aligned} \mathcal{L} &:= \{\text{User}, \text{Trip}, \text{PlaceEvent}, \text{Place}\} \\ \sigma(\text{User}) &:= \sigma(\text{Place}) := 1 \\ \sigma(\text{PlaceEvent}) &:= \text{Place} \times \text{Integer} \\ \sigma(\text{Trip}) &:= \text{User} \times \text{User} \\ &\quad \times (1 + \text{PlaceEvent}) \times (1 + \text{PlaceEvent}) \end{aligned}$$

Given an APG schema $S = (\mathcal{L}, \sigma)$, an APG on S consists of:

- ▶ $E : \mathcal{L} \rightarrow \mathbf{Set}$
- ▶ For each $l \in \mathcal{L}$, a function $v_l : E(l) \rightarrow V(\sigma(l))$, where $V : \mathcal{T} \rightarrow \mathbf{Set}$ is defined as follows:

$$\begin{aligned} V(0) &:= 0 \quad V(1) := 1 \\ V(\text{Prim } p) &:= PV(p) \quad V(\text{Lbl } l) := E(l) \\ V(t_1 + t_2) &:= V(t_1) + V(t_2) \\ V(t_1 \times t_2) &:= V(t_1) \times V(t_2) \end{aligned}$$

Example APG on S :

$$\begin{aligned} E(\text{User}) &:= \{u_1, u_2, u_3\} \\ E(\text{Trip}) &:= \{t_1, t_2\} \\ E(\text{PlaceEvent}) &:= \{e_1, e_2, e_3\} \\ E(\text{Place}) &:= \{p_1, p_2, p_3\} \\ v_{\text{User}}(u_1) &:= v_{\text{User}}(u_2) := v_{\text{User}}(u_3) := () \\ v_{\text{Place}}(p_1) &:= v_{\text{Place}}(p_2) := v_{\text{Place}}(p_3) := () \\ v_{\text{Trip}}(t_1) &:= (u_1, u_2, \text{inr}(e_1), \text{inr}(e_2)) \\ v_{\text{Trip}}(t_2) &:= (u_1, u_3, \text{inr}(e_3), \text{inl}(())) \\ v_{\text{PlaceEvent}}(e_1) &:= (p_1, 1602203601) \\ v_{\text{PlaceEvent}}(e_2) &:= (p_2, 1602203948) \\ v_{\text{PlaceEvent}}(e_3) &:= (p_2, 1602204122) \end{aligned}$$

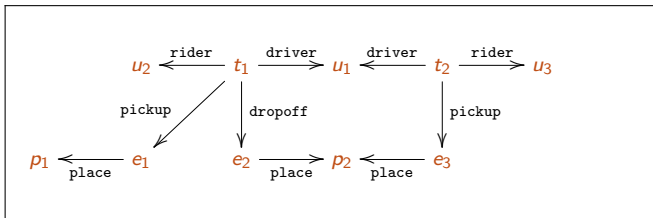
APGs: Definition 2 (type-theoretic flavor)

Fix a set P of primitive types p with extensions $PV(p)$.

An APG schema S consists of:

- ▶ $\mathcal{L} : \text{Set}$
- ▶ $\sigma : \mathcal{L} \rightarrow \mathcal{T}$, where

$$\mathcal{T} \ni t ::= 0 \mid 1 \mid t_1 + t_2 \mid t_1 \times t_2 \mid \text{Prim } p \ (p \in P) \\ \mid \text{Lbl } l \ (l \in \mathcal{L})$$



Let $P := \{\text{Integer}\}$, $PV(\text{Integer}) = \mathbb{Z}$.

Example schema $S = (\mathcal{L}, \sigma)$:

$$\begin{aligned} \mathcal{L} &:= \{\text{User}, \text{Trip}, \text{PlaceEvent}, \text{Place}\} \\ \sigma(\text{User}) &:= \sigma(\text{Place}) := 1 \\ \sigma(\text{PlaceEvent}) &:= \text{Place} \times \text{Integer} \\ \sigma(\text{Trip}) &:= \text{User} \times \text{User} \\ &\quad \times (1 + \text{PlaceEvent}) \times (1 + \text{PlaceEvent}) \end{aligned}$$

Given an APG schema $S = (\mathcal{L}, \sigma)$, an APG on S consists of:

- ▶ $\mathcal{E} : \text{Set}$
- ▶ $\lambda : \mathcal{E} \rightarrow \mathcal{L}$
- ▶ $v : \mathcal{E} \rightarrow \mathcal{V}$ where

$$\mathcal{V} \ni (v : t) ::= () \mid 1 \mid \text{inl}_{t_2}(v : t_1) : t_1 + t_2 \\ \mid \text{inr}_{t_1}(v : t_2) : t_1 + t_2 \\ \mid (v_1 : t_1, v_2 : t_2) : t_1 \times t_2 \\ \mid \text{Prim } v_t : t \ (t \in P, v \in PV(p)) \\ \mid \text{Elmt } e : \lambda(e) \ (e \in \mathcal{E})$$

such that, where $\tau : \mathcal{V} \rightarrow \mathcal{T}$ is defined by $\tau(v : t) := t$,

$$\begin{array}{ccc} \mathcal{E} & \xrightarrow{\lambda} & \mathcal{L} \\ v \downarrow & & \downarrow \sigma \\ \mathcal{V} & \xrightarrow{\tau} & \mathcal{T} \end{array}$$

Example APG on S :

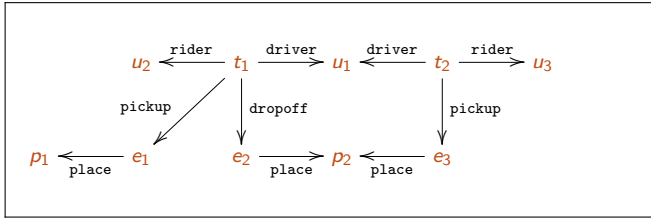
$$\begin{aligned} \mathcal{E} &:= \{u_1, u_2, u_3, t_1, t_2, e_1, e_2, e_3, p_1, p_2, p_3\} \\ \lambda(u_1) &:= \lambda(u_2) := \lambda(u_3) := \text{User} \quad \lambda(t_1) := \lambda(t_2) := \text{Trip} \\ \lambda(p_1) &:= \lambda(p_2) := \lambda(p_3) := \text{Place} \\ \lambda(e_1) &:= \lambda(e_2) := \lambda(e_3) := \text{PlaceEvent} \\ v(u_1) &:= v(u_2) := v(u_3) := v(p_1) := v(p_2) := v(p_3) := () \\ v(t_1) &:= (u_1, u_2, \text{inr}(e_1), \text{inr}(e_2)) \\ v(t_2) &:= (u_1, u_3, \text{inr}(e_3), \text{inl}(())) \\ v(e_1) &:= (p_1, 1602203601) \quad v(e_2) := (p_2, 1602203948) \\ v(e_3) &:= (p_2, 1602204122) \end{aligned}$$

APGs: Definition 3: APGs as Coalgebras

A APG schema S consists of:

- ▶ $\mathcal{L} : \text{Set}$
- ▶ a polynomial functor $F : \text{Set}^{\mathcal{L}} \rightarrow \text{Set}^{\mathcal{L}}$

(This definition includes non-computable schemas, but it is mathematically very nice.)



Example schema $S = (\mathcal{L}, F)$:

$\mathcal{L} := \{\text{User}, \text{Trip}, \text{PlaceEvent}, \text{Place}\}$

$$F \left(\begin{bmatrix} \text{User} \mapsto E_{\text{User}} \\ \text{Trip} \mapsto E_{\text{Trip}} \\ \text{PlaceEvent} \mapsto E_{\text{PlaceEvent}} \\ \text{Place} \mapsto E_{\text{Place}} \end{bmatrix} \right) :=$$

$$\begin{bmatrix} \text{User} \mapsto 1 \\ \text{Trip} \mapsto 1 \\ \text{PlaceEvent} \mapsto E_{\text{Place}} \times \mathbb{Z} \\ \text{Place} \mapsto E_{\text{User}} \times E_{\text{User}} \times (1 + E_{\text{PlaceEvent}}) \times (1 + E_{\text{PlaceEvent}}) \end{bmatrix}$$

Given an APG schema $S = (\mathcal{L}, F)$, an APG on S consists of:

- ▶ $E : \mathcal{L} \rightarrow \text{Set}$
- ▶ $v : E \Rightarrow F(E)$

In other words, it is a coalgebra of F .

(This definition includes non-computable APGs, but it is mathematically very nice.)

Now we can define morphisms of APGs on S simply as morphisms of coalgebras of F .

Example APG on S :

$$\begin{aligned} E(\text{User}) &:= \{u_1, u_2, u_3\} \\ E(\text{Trip}) &:= \{t_1, t_2\} \\ E(\text{PlaceEvent}) &:= \{e_1, e_2, e_3\} \\ E(\text{Place}) &:= \{p_1, p_2, p_3\} \\ v_{\text{User}}(u_1) &:= v_{\text{User}}(u_2) := v_{\text{User}}(u_3) := () \\ v_{\text{Place}}(p_1) &:= v_{\text{Place}}(p_2) := v_{\text{Place}}(p_3) := () \\ v_{\text{Trip}}(t_1) &:= (u_1, u_2, \text{inr}(e_1), \text{inr}(e_2)) \\ v_{\text{Trip}}(t_2) &:= (u_1, u_3, \text{inr}(e_3), \text{inl}(())) \\ v_{\text{PlaceEvent}}(e_1) &:= (p_1, 1602203601) \\ v_{\text{PlaceEvent}}(e_2) &:= (p_2, 1602203948) \\ v_{\text{PlaceEvent}}(e_3) &:= (p_2, 1602204122) \end{aligned}$$

APGs: Definition 4: APGs as models

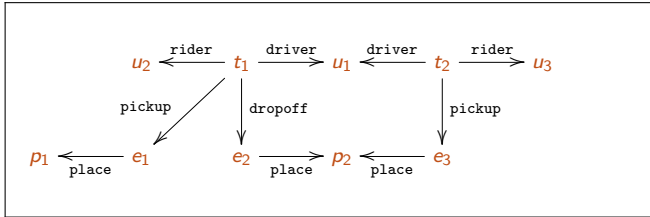
Let $S = (\mathcal{L} : \mathbf{Set}, \sigma : \mathcal{L} \rightarrow \mathcal{T})$ be an APG schema.

The APG theory C_S corresponding to S is the free category C_S with terminal and initial object, products, and coproducts, on the following generators:²

$$\frac{l \in \mathcal{L}}{(\text{Lbl } l) \in C_S} \quad \frac{p \in P}{(\text{Prim } p) \in C_S} \quad \frac{l \in \mathcal{L}}{\delta_l : l \rightarrow \sigma(l)}$$

Note that the objects of C_S are exactly \mathcal{T} .

(We can also extend APGs by adding more generators to C_S , e.g. user-defined functions and constraints.)



Let $P := \{\text{Integer}\}$, $PV(\text{Integer}) = \mathbb{Z}$.

Example schema $S = (\mathcal{L}, \sigma)$:

$$\begin{aligned} \mathcal{L} &:= \{\text{User}, \text{Trip}, \text{PlaceEvent}, \text{Place}\} \\ \sigma(\text{User}) &:= \sigma(\text{Place}) := 1 \\ \sigma(\text{PlaceEvent}) &:= \text{Place} \times \text{Integer} \\ \sigma(\text{Trip}) &:= \text{User} \times \text{User} \\ &\quad \times (1 + \text{PlaceEvent}) \times (1 + \text{PlaceEvent}) \end{aligned}$$

Given an APG schema C_S , a APG on S is a model of C_S , that is, a functor $V : C_S \rightarrow \mathbf{Set}$ which preserves products, coproducts, terminal object, and initial object, and which sends $\text{Prim } p \mapsto PV(p)$.

Note that V is determined completely by where it sends objects of the form $\text{Lbl } l$ and morphisms of the form δ_l .

(This definition includes non-computable APGs, but it is mathematically very nice.)

Now we can define morphisms of APGs V and V' on S as natural transformations $\eta : V \Rightarrow V'$ such that

$$\eta_{\text{Prim } p} = \text{id}_{\text{Prim } p} \quad \eta_{t+t'} = \eta_t + \eta_{t'} \quad \eta_{t \times t'} = \eta_t \times \eta_{t'}$$

Example APG on S :

$$V(\text{Lbl User}) := \{u_1, u_2, u_3\}$$

$$V(\text{Lbl Trip}) := \{t_1, t_2\}$$

$$V(\text{Lbl PlaceEvent}) := \{e_1, e_2, e_3\}$$

$$V(\text{Lbl Place}) := \{p_1, p_2, p_3\}$$

$$V(\delta_{\text{User}}) := \begin{bmatrix} u_1 \mapsto () \\ u_2 \mapsto () \\ u_3 \mapsto () \end{bmatrix} \quad V(\delta_{\text{Place}}) := \begin{bmatrix} p_1 \mapsto () \\ p_2 \mapsto () \\ p_3 \mapsto () \end{bmatrix}$$

$$V(\delta_{\text{Trip}}) := \begin{bmatrix} t_1 \mapsto (u_1, u_2, \text{inr}(e_1), \text{inr}(e_2)) \\ t_2 \mapsto (u_1, u_3, \text{inr}(e_3), \text{inl}(())) \end{bmatrix}$$

$$V(\delta_{\text{PlaceEvent}}) := \begin{bmatrix} e_1 \mapsto (p_1, 1602203601) \\ e_2 \mapsto (p_2, 1602203948) \\ e_3 \mapsto (p_2, 1602204122) \end{bmatrix}$$

²See Cockett and Seely 2001 for the details of this construction. We also could choose to require that in C_S , the canonical maps $t \times u + t \times v \rightarrow t \times (u + v)$ and $t \rightarrow t + 0$ are isomorphisms, making C_S a distributive category. This has no effect on the models of C_S , since \mathbf{Set} is distributive, but it does afford additional morphisms of theories (see next slide). The downside is that it complicates term rewriting.

Morphisms of APG Theories and Schemas

Morphisms of APG theories

A morphism of APG theories C_S and $C_{S'}$ is defined as a functor $\Phi : C_S \rightarrow C_{S'}$ preserving products, coproducts, terminal object, initial object, and primitive types.

Now we have a functor $\Delta : \mathbf{Theory}^{\text{op}} \rightarrow \mathbf{Cat}$, where **Theory** is the category of APG theories, defined by:

- ▶ $\Delta(C_S) := S\text{-APG}$, where $S\text{-APG}$ is the category of APGs on S
- ▶ $\Delta(\Phi) := (V \mapsto V \circ \Phi)$

Simple example

Source schema $S = (\mathcal{L}, \sigma)$:

- ▶ $\mathcal{L} := \{l\}$, $\sigma(l) := \text{String} \times \text{Nat} \times \text{Integer}$

Target schema $S' = (\mathcal{L}', \sigma')$:

- ▶ $\mathcal{L} := \{l'\}$, $\sigma'(l') := \text{Nat} \times \text{String}$

Theory mapping $\Phi : C_S \rightarrow C_{S'}$:

- ▶ $\Phi(\text{Lbl } l) := \text{Lbl } l'$
- ▶ $\Phi(\delta_l) := l' \xrightarrow{\delta_{l'}} \text{Nat} \times \text{String} \xrightarrow{\langle \text{snd}, \text{fst}, 42 \circ ! \rangle} \text{String} \times \text{Nat} \times \text{Integer}$

(Here we assume a user-defined function $42 : 1 \rightarrow \text{String}$ sending $() \mapsto 42$).

The functor $\Delta(\Phi)$ converts APGs on S' to schema S by permuting projections and adding 42 .

Morphisms of APG schemas

A morphism of schemas (\mathcal{L}, F) and (\mathcal{L}', F') is a pair $(\Phi : \text{Set}^{\mathcal{L}'} \rightarrow \text{Set}^{\mathcal{L}}$ polynomial functor, $\phi : \Phi F_{S'} \Rightarrow F_S \Phi)$, as shown:

$$\begin{array}{ccc} \text{Set}^{\mathcal{L}'} & \xrightarrow{F_{S'}} & \text{Set}^{\mathcal{L}'} \\ \downarrow \Phi & \swarrow \phi & \downarrow \Phi \\ \text{Set}^{\mathcal{L}} & \xrightarrow{F_S} & \text{Set}^{\mathcal{L}} \end{array}$$

Now we have a functor $\Delta : \mathbf{Schema}^{\text{op}} \rightarrow \mathbf{Cat}$, where **Schema** is the category of APG schemas, defined by:

- ▶ $\Delta(C_S) := S\text{-APG}$, where $S\text{-APG}$ is the category of APGs on S
- ▶ $\Delta(\Phi, \phi) := ((E' : \mathcal{L}' \rightarrow \mathbf{Set}, \mathbf{v}' : E' \Rightarrow F'(E')) \mapsto (\Phi(E'), (\Phi(E') \xrightarrow{\Phi(\mathbf{v}')} \Phi F'(E') \xrightarrow{\phi_{E'}} F\Phi(E')))$

These morphisms can be translated to morphisms of the corresponding theories by a functor $C : \mathbf{Schema} \rightarrow \mathbf{Theory}$ sending $S \mapsto C_S$ and satisfying

$$\begin{array}{ccc} \mathbf{Schema}^{\text{op}} & \xrightarrow{C} & \mathbf{Theory}^{\text{op}} \\ & \searrow \Delta & \swarrow \Delta \\ & \mathbf{Cat} & \end{array}$$

Though faithful, C is not full, making morphisms of APG theories strictly more general than morphisms of APG schemas!

In the paper on arXiv, there is a conjecture on how morphisms of APG schemas can be generalized in order to extend this functor to an equivalence.

Stream Example

Streams can be considered as APGs on the schema $S = (\mathcal{L}, \sigma)$:

- ▶ $\mathcal{L} := \{A\}$
- ▶ $\sigma(A) := A + 1$

Finite stream:

- ▶ $E(A) := \{e_1, e_2, e_3, \dots, e_{10}\}$
- ▶ $v(e_n) := \text{inl}(e_{n+1})$ for $n = 1, \dots, 9$
- ▶ $v(e_{10}) := \text{inr}()$

Infinite stream:

- ▶ $E(A) := \{e_1, e_2, e_3, \dots, e_{10}\}$
- ▶ $v(e_n) := \text{inl}(e_{n+1})$ for $n = 1, \dots, 9$
- ▶ $v(e_{10}) := \text{inl}(e_4)$

A morphism from the theory C_S to itself (assume C_S is taken as distributive).

$$\Phi(A) := A \times A$$

$$\Phi(\delta_A) : A \times A \rightarrow A \times A + 1$$

$$\Phi(\delta_A) := \left(A \times A \xrightarrow{\delta_A \times \delta_A} (A + 1) \times (A + 1) \right.$$

$$\left. \xrightarrow{(\delta_{A+1}) \times \text{id}} ((A + 1) + 1) \times (A + 1) \right.$$

$$\left. \xrightarrow{\text{dist, assoc}} A \times A + (1 + 1) \times A + A \times 1 + (1 + 1) \times 1 \right.$$

$$\left. \xrightarrow{[\text{id}, !, !]} A \times A + 1 \right)$$

We apply this morphism to the finite stream (E, v) to obtain the stream $\Delta(\Phi)(E, v) =: (E', v')$:

$$E'(A) := \{(e_i, e_j) \mid i, j = 1, \dots, 10\}$$

$$v'((e_i, e_j)) := \text{inl}((e_{i+2}, e_{j+1})) \text{ for } i = 1, \dots, 8 \text{ and } j = 1, \dots, 9$$

$$v'((e_i, e_{10})) := \text{inr}() \text{ for } i = 1, \dots, 10$$

$$v'((e_i, e_j)) := \text{inr}() \text{ for } i = 9, 10 \text{ and } j = 1, \dots, 10$$

We can say that $\Delta(\Phi)$ converts a stream into a new stream whose elements are ordered pairs of elements of the old stream, and where the first element advances twice as fast as in the old stream, and the second element advances just as in the old stream.

Generalized APGs (GAPGs)

In the language of David Spivak (cf. *The Poly Book*), polynomial functors $\mathbf{Set}^{\mathcal{L}} \rightarrow \mathbf{Set}^{\mathcal{L}}$ can be described as bicomodules in Poly from a discrete category (aka comonoid) to itself: $\mathcal{L}y \xleftarrow{\sigma} \triangleleft \mathcal{L}y$.

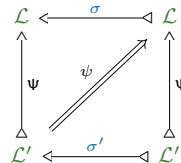
We immediately generalize to the case of an arbitrary category \mathcal{L} of labels, to obtain a *generalized APG (GAPG) schema*, defined as a category \mathcal{L} and a bicomodule (aka prafunctor) $\mathcal{L} \xleftarrow{\sigma} \triangleleft \mathcal{L}$. A GAPG on the schema (\mathcal{L}, σ) is then an ordered pair (E, ν) where E is a copresheaf on \mathcal{L} , aka a bicomodule $\mathcal{L} \xleftarrow{E} \triangleleft 0$, and ν is a bicomodule homomorphism:



Use cases and implementation possibilities of GAPGs have yet to be explored.

Morphisms of GAPG schemas

A schema morphism $(\mathcal{L}, \sigma) \rightarrow (\mathcal{L}', \sigma')$ is an ordered pair (Ψ, ψ) :



Then the schema morphism (Ψ, ψ) can be applied to a GAPG (E', ν') on (\mathcal{L}', σ') in a straightforward way:

